Automating Machine Learning for Prevention Research

Jason H. Moore, PhD, FACMI

Edward Rose Professor of Informatics
Director, Institute for Biomedical Informatics
Senior Associate Dean for Informatics

Perelman School of Medicine
University of Pennsylvania
Philadelphia, PA, USA

upibi.org epistasis.org jhmoore@upenn.edu

Biomedical Informatics

Basic Science	• Bioinformatics • Computational Biology
Clinical	• Clinical Informatics • Clinical Research Informatics • Consumer Health Informatics
Population	• Public Health Informatics

American Medical Informatics Association (AMIA)

Golden Era of Biomedical Informatics

Moore and Holmes, *BioData Mining* (2016)

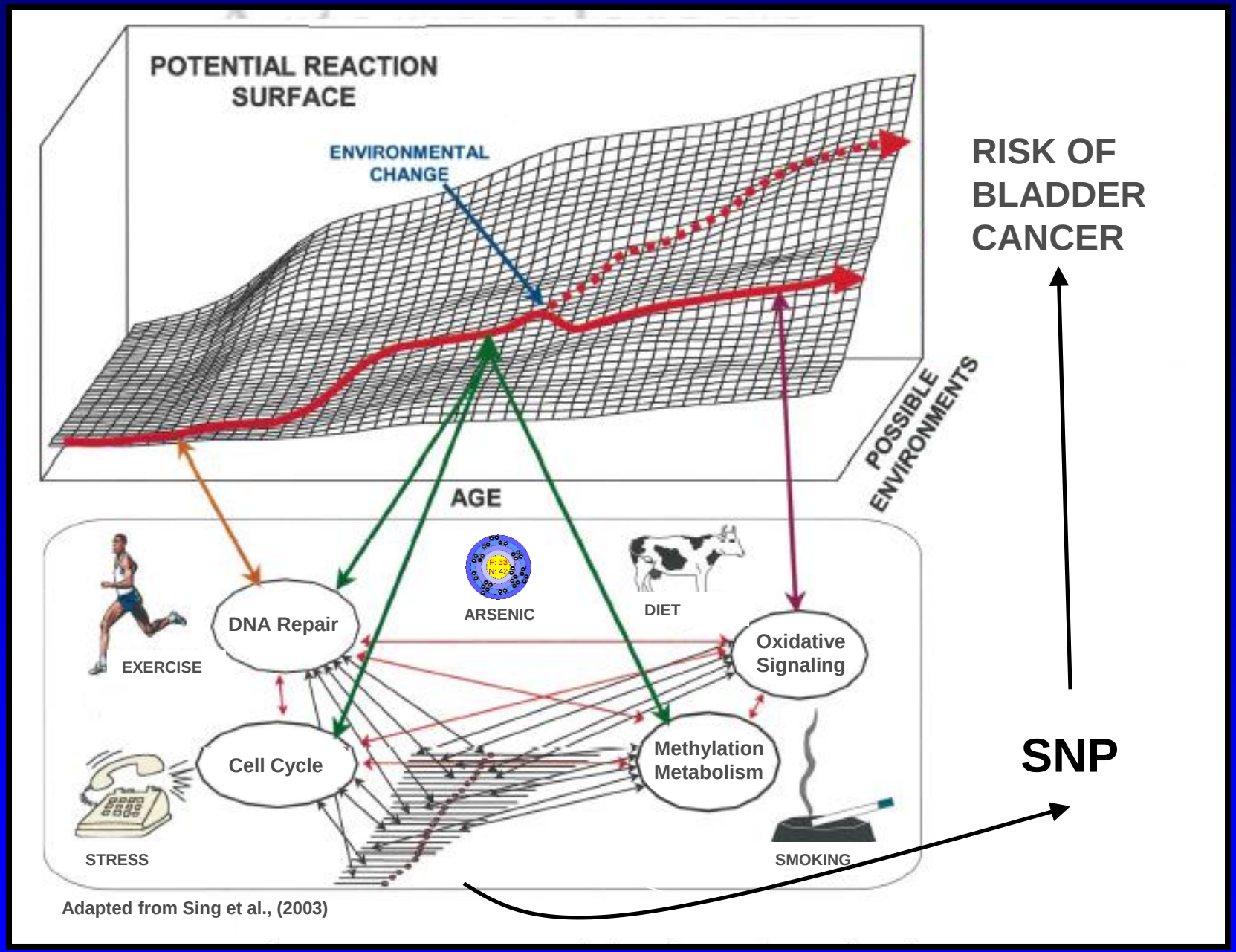
Why?

- Big data
- High-performance computing
- Talented trainees
- Government recognition
- Industry recognition
- Patient recognition
- University investment

Golden Era of Biomedical Informatics

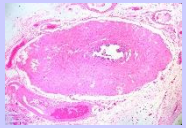
What next?

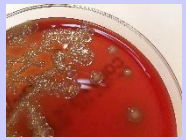
- Artificial intelligence
- Biomedical devices
- Data integration
- Data science
- Informatician scientists
- Machine Learning
- No-boundary thinking
- Visual analytics


POTENTIAL REACTION SURFACE
ENVIRONMENTAL CHANGE
RISK OF BLADDER CANCER
POSSIBLE ENVIRONMENTS
AGE
EXERCISE
DNA Repair
ARSENIC
DIET
Oxidative Signaling
Cell Cycle
Methylation Metabolism
STRESS
SMOKING
SNP
Adapted from Sing et al., (2003)

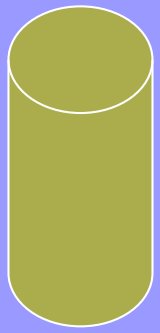
Data Science Pipeline

D





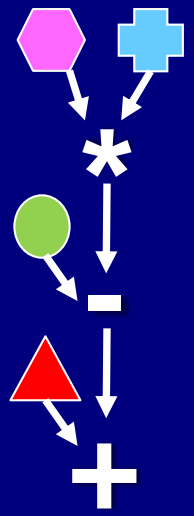
DI

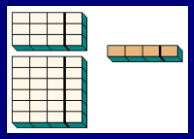


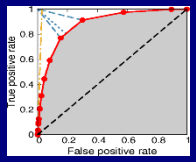
FS

FC

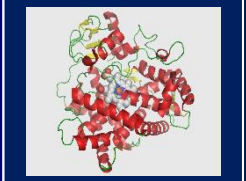
ML

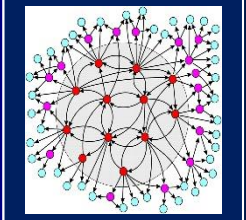


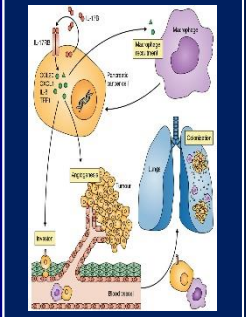




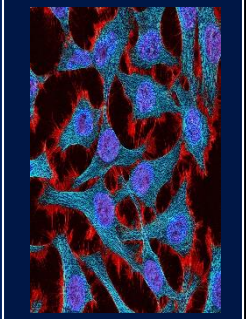
I







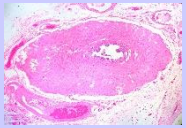
V

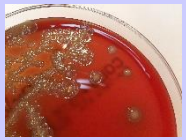


A

Big Data

D



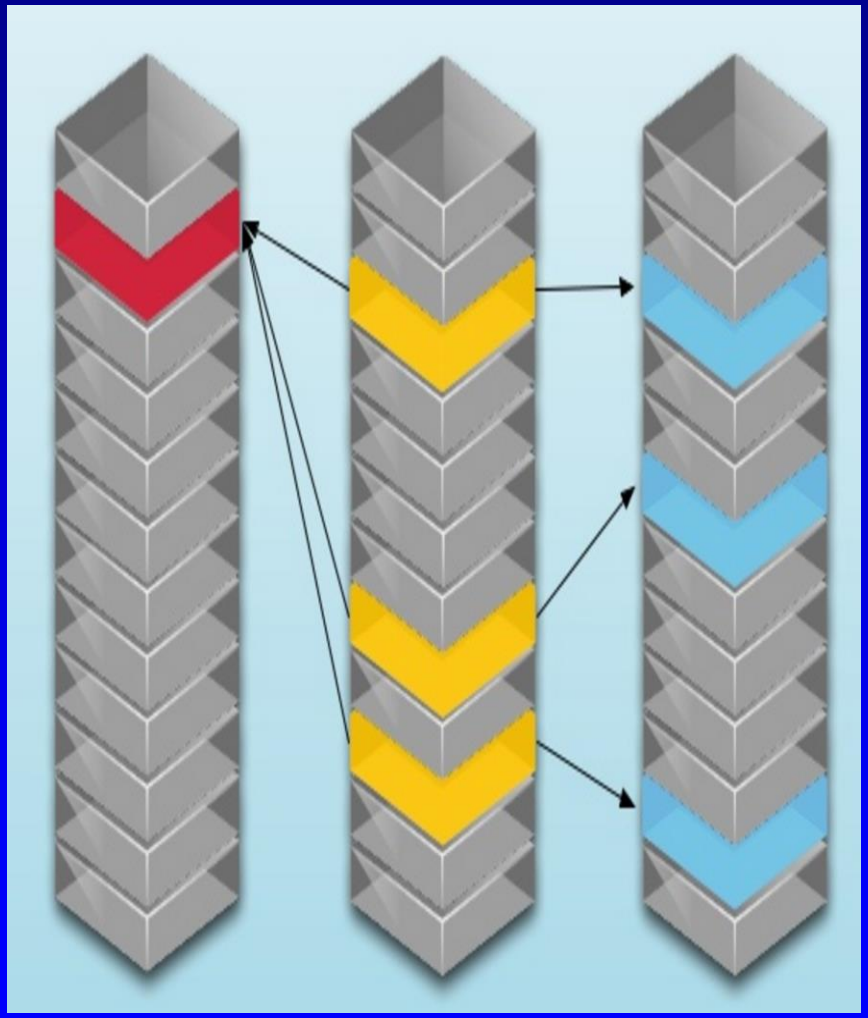


http://www.kdnuggets.com/

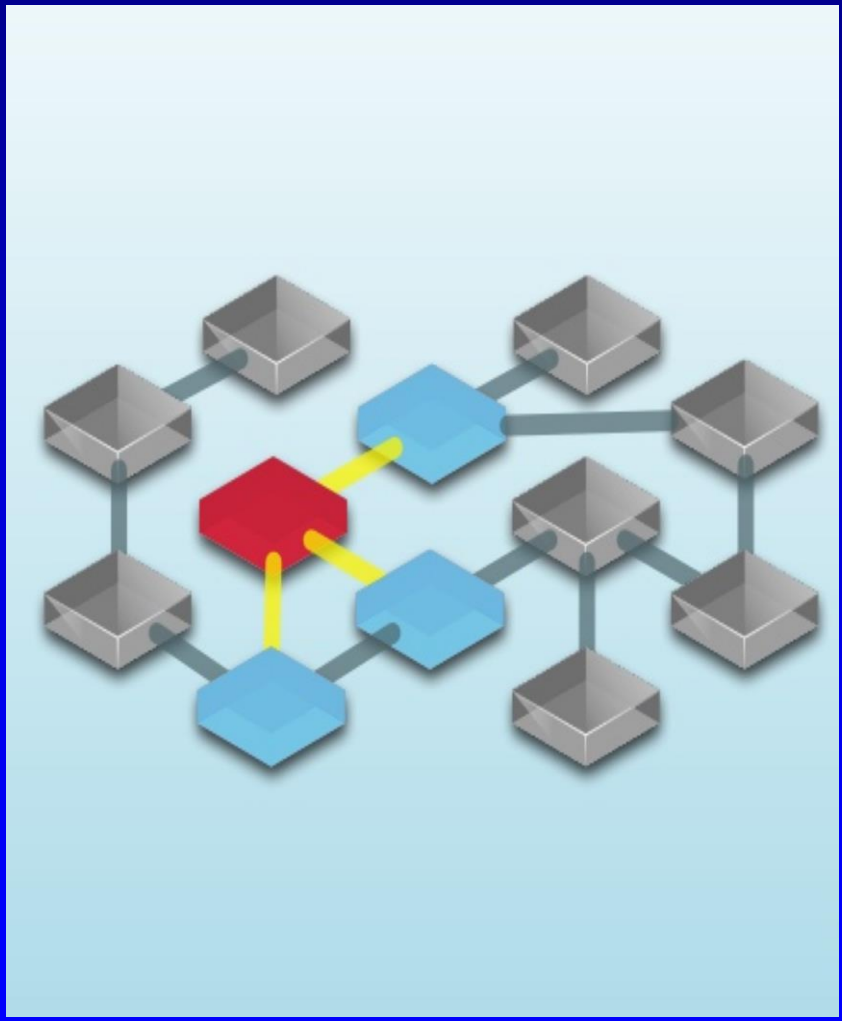
Data Integration

DI

Relational Database

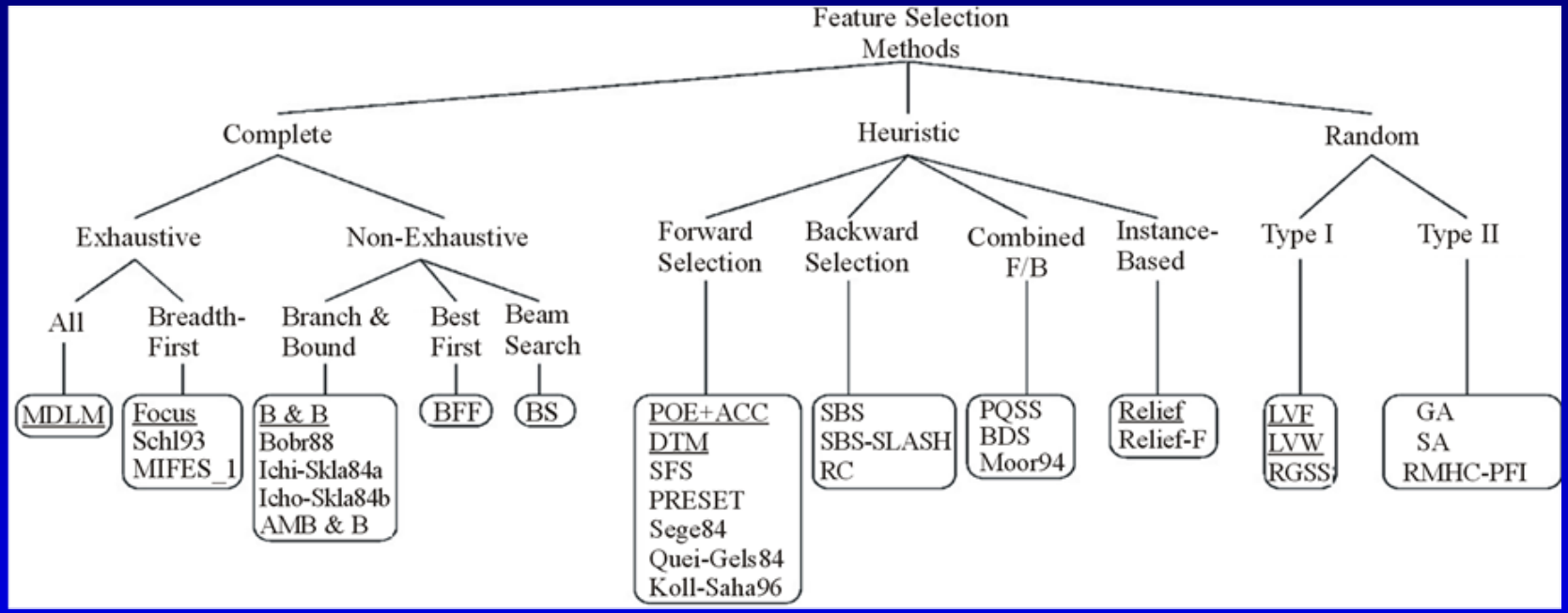


Graph Database

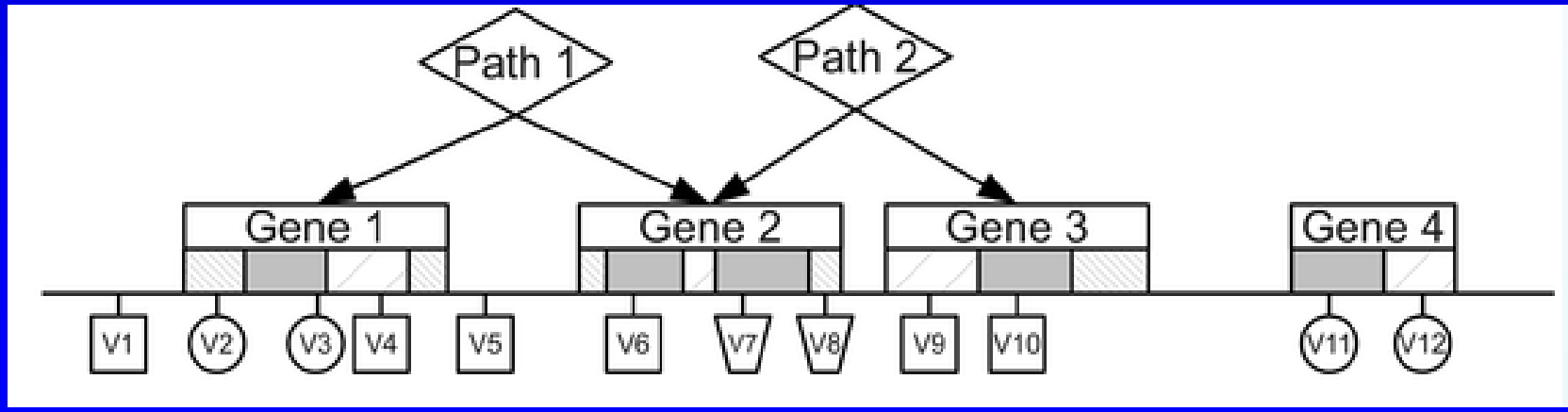


Feature Selection

FS

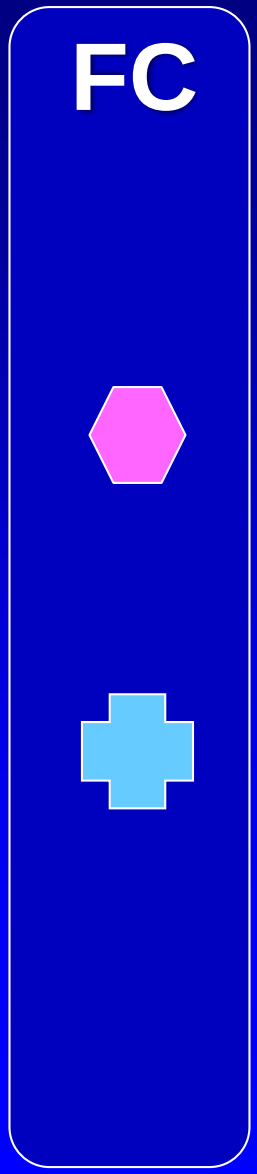


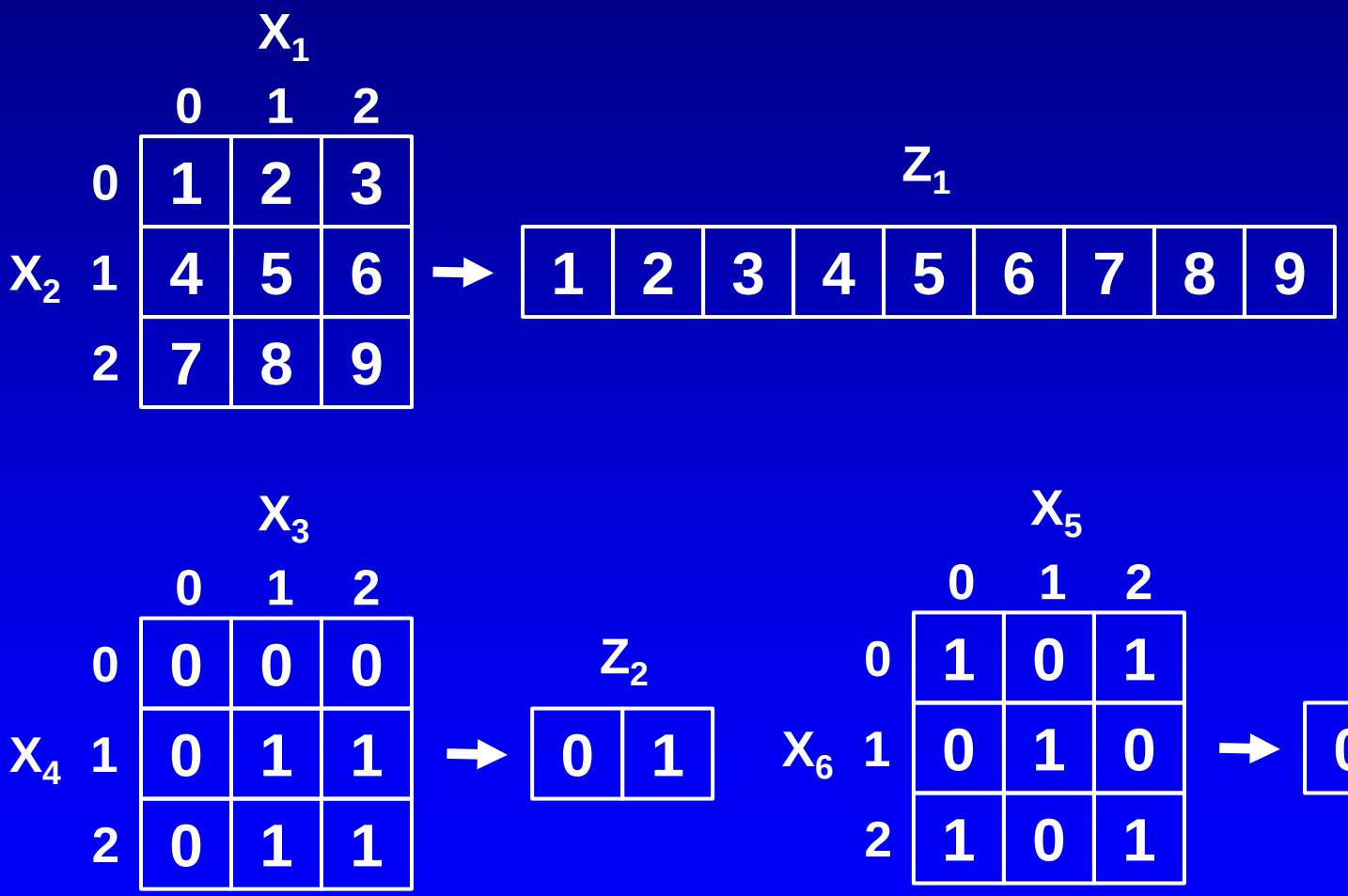
Sohangir – J Soft Engin App (2013)



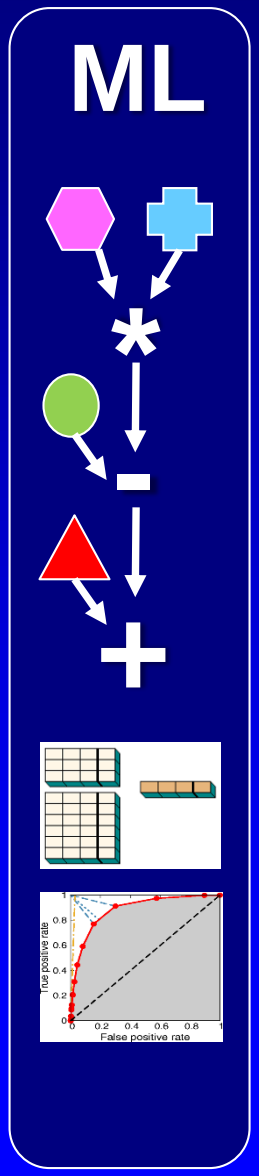
Ritchie – PLoS Genetics (2013)

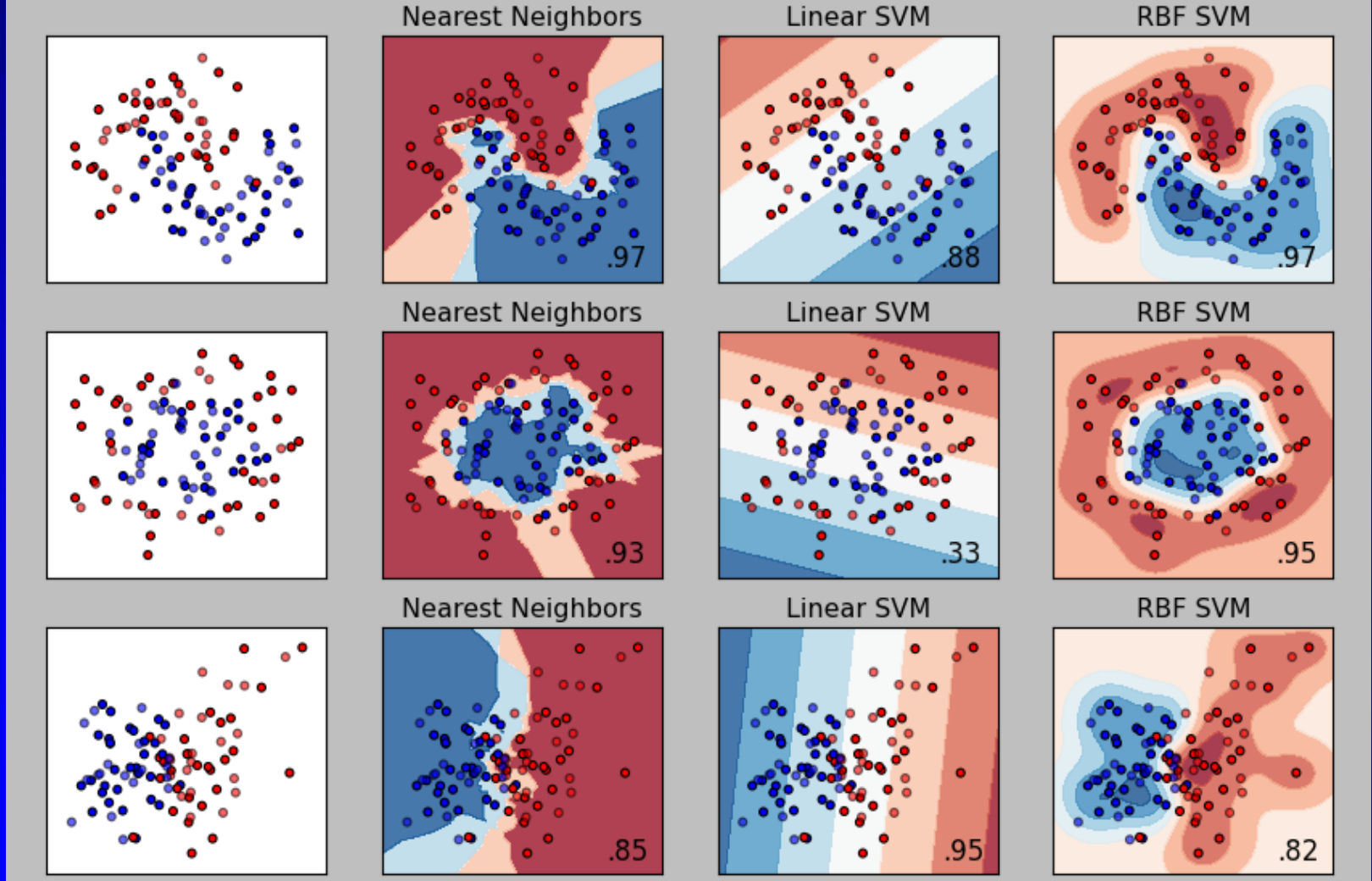
Feature Construction





Machine Learning

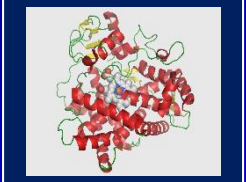


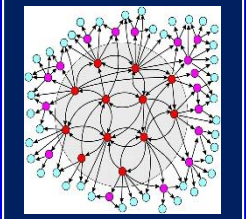


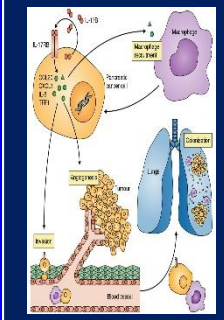
http://suanfazu.com/

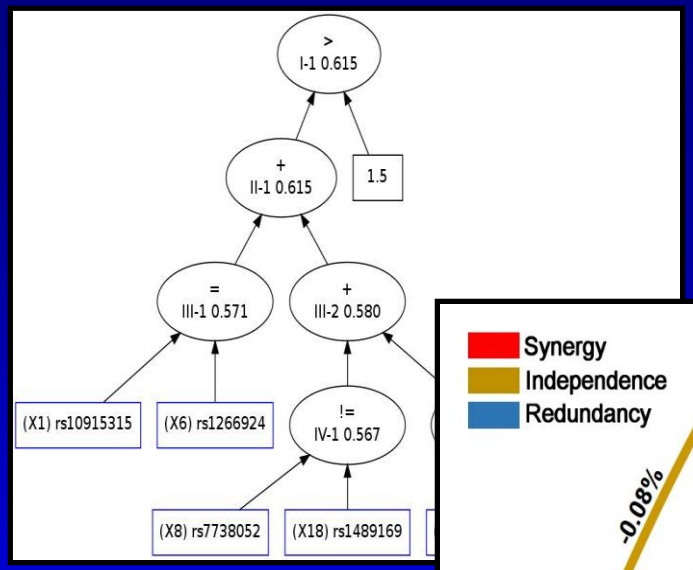
Statistical and Biological Interpretation

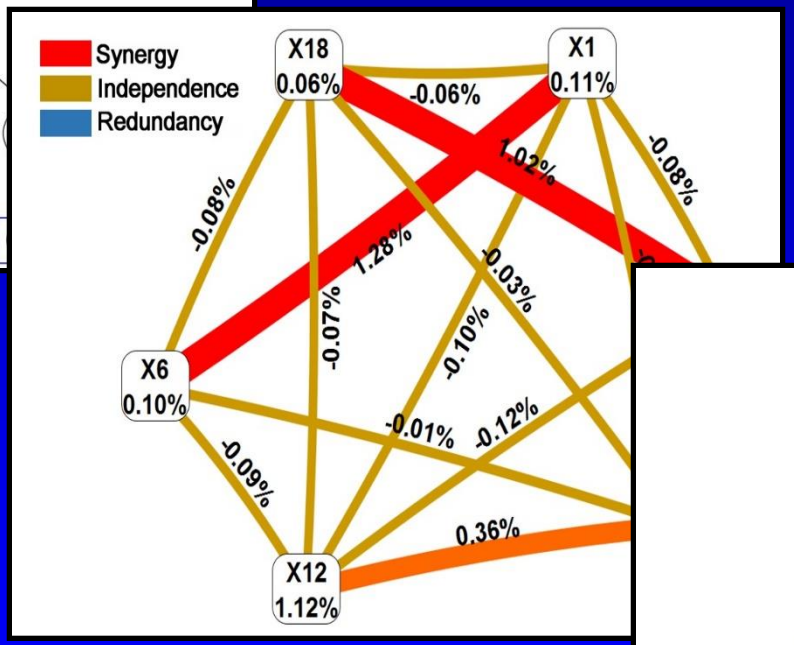
I

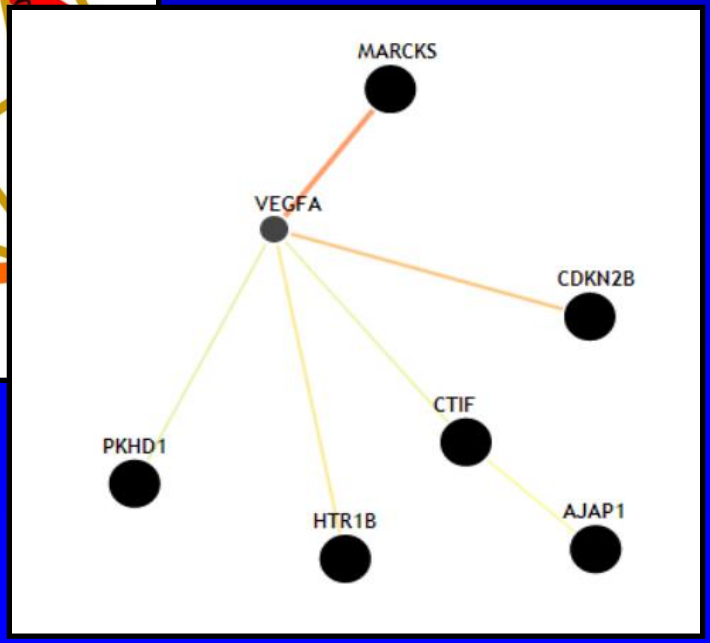






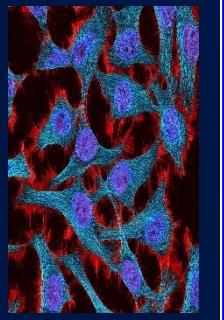


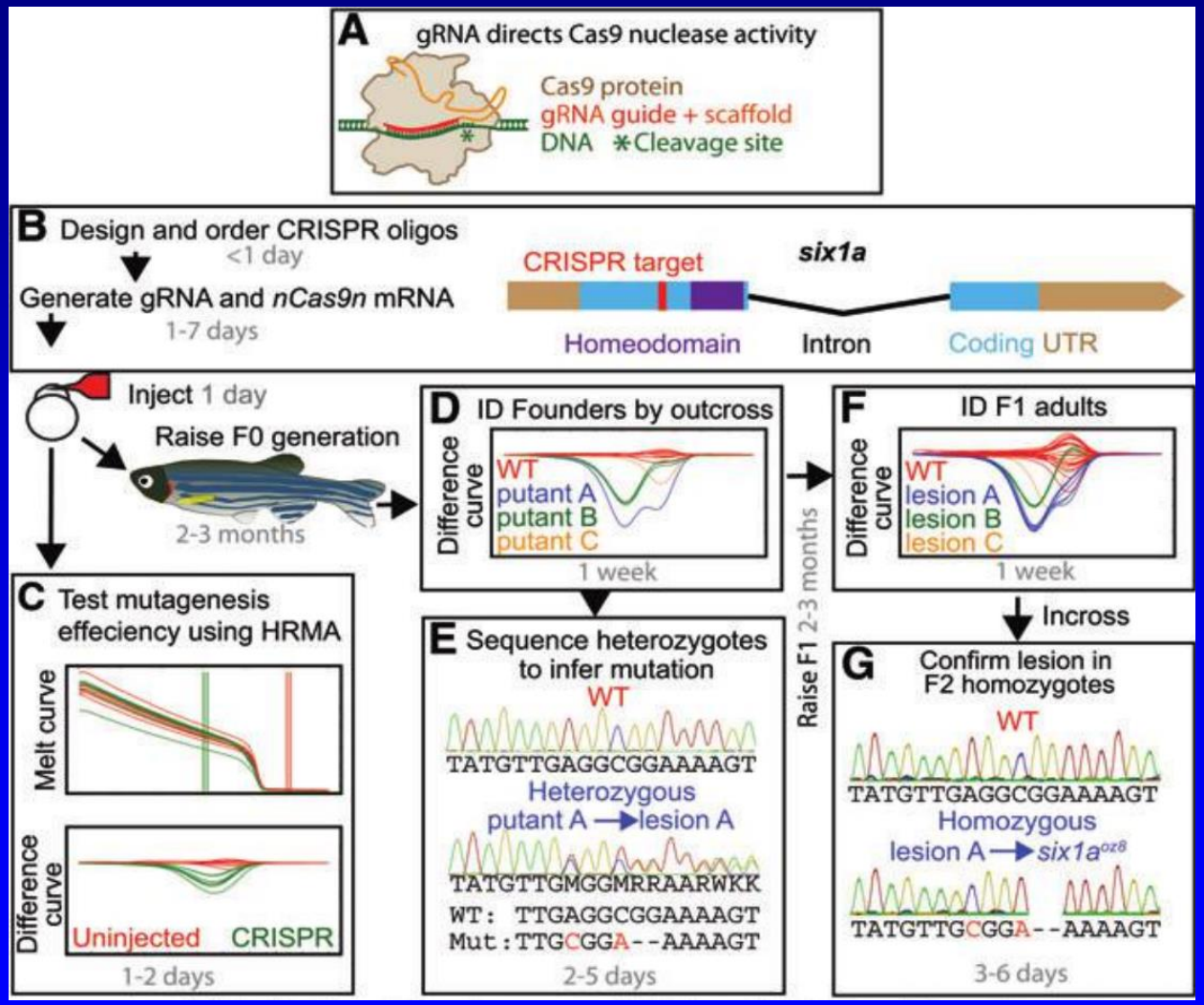




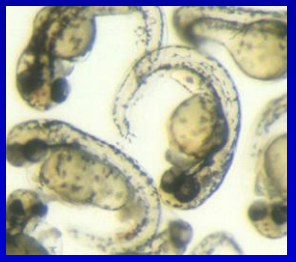
Biological Validation

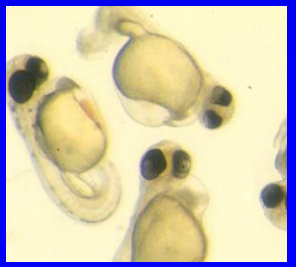
V





Talbot, Zebrafish (2014)

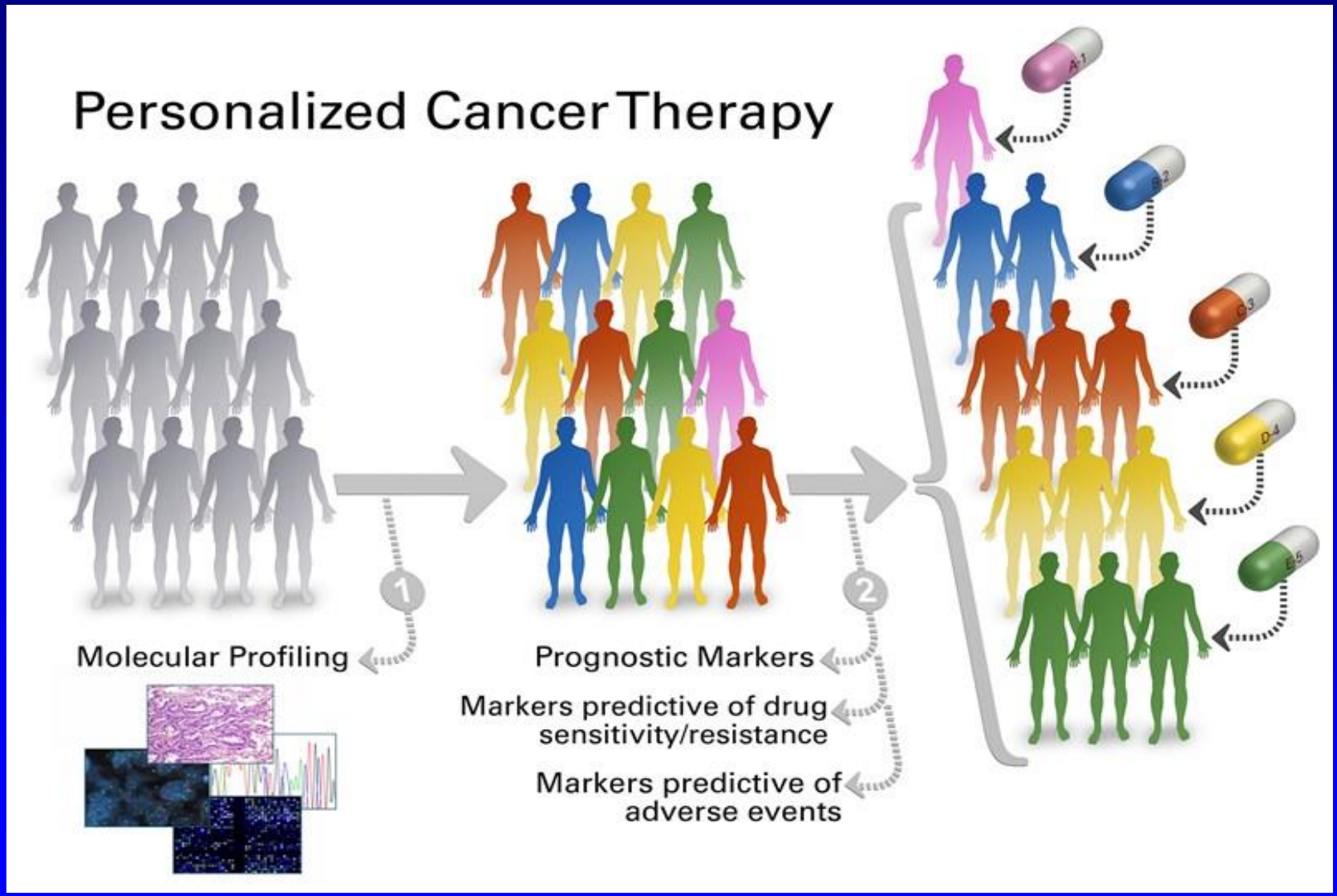




dev.biologists.org

Clinical Application

A

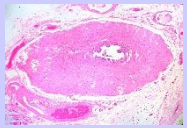


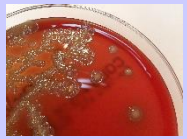
M.D. Anderson

Can the data science pipeline construction process be automated?

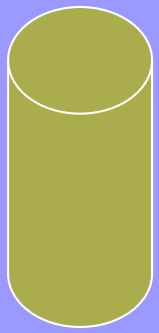
Automated Machine Learning (AutoML)

D





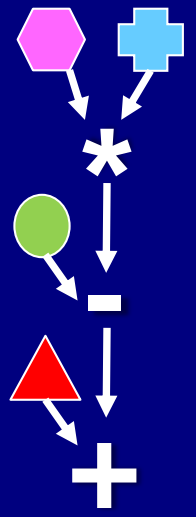
DI

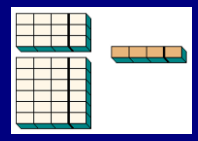


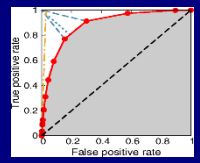
FS

FC

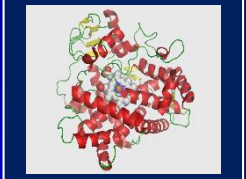
ML

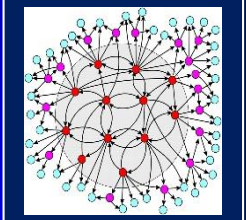


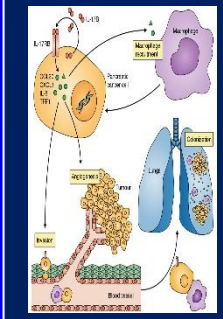




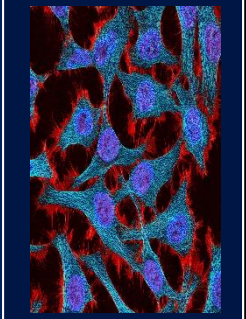
I



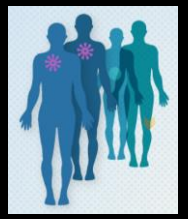




V



A



Tree-Based Pipeline Optimization Tool (TPOT)

Towards Automated Data Science

https://github.com/epistasislab/tpot/

Dr. Randal Olson (postdoc)

Tree-Based Pipeline Optimization Tool (TPOT)

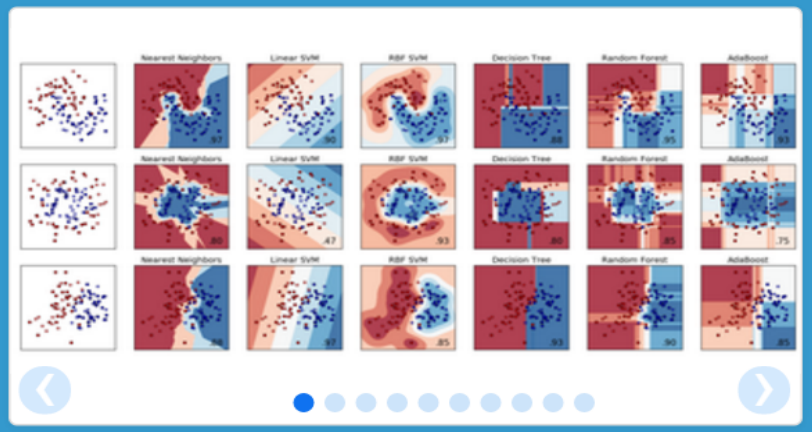
1) ML code base

2) Pipeline representation

3) Optimization algorithm

4) Overfitting control

ML code base



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

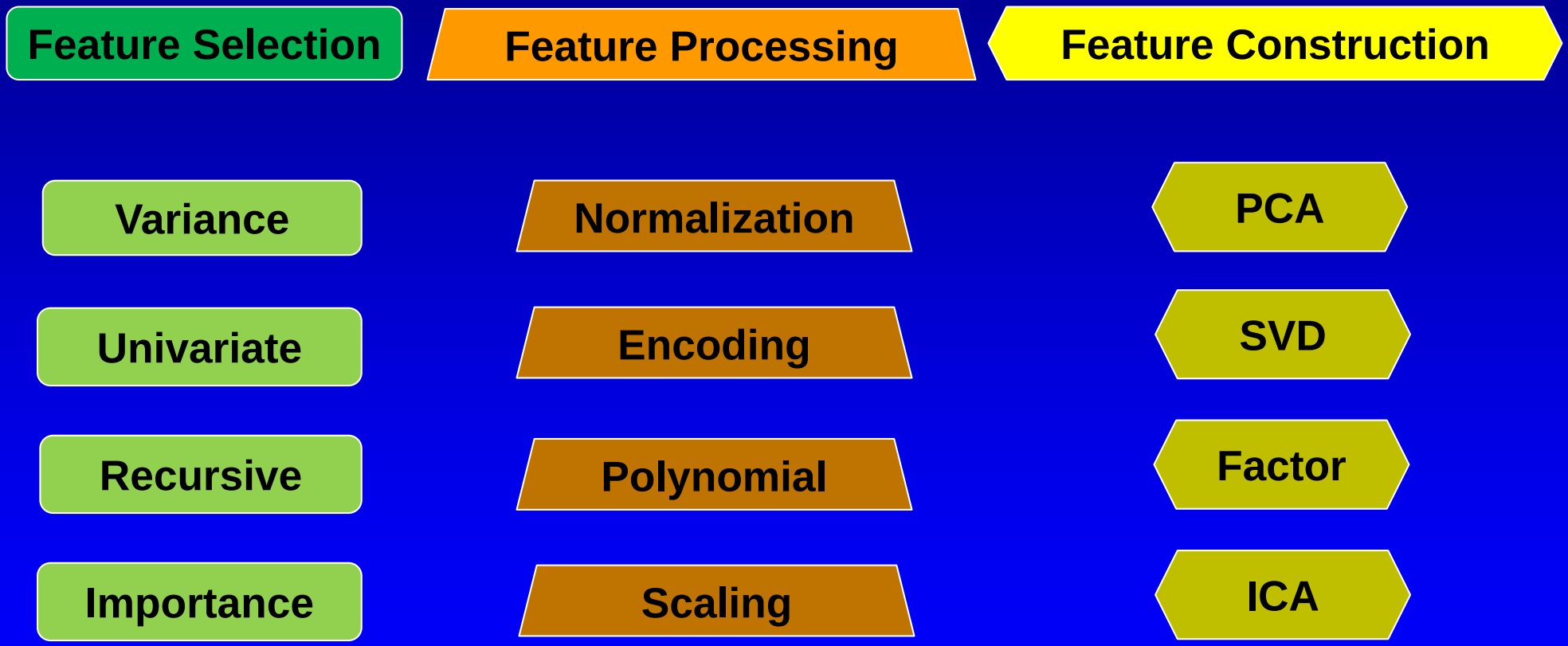
Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

Preprocessing

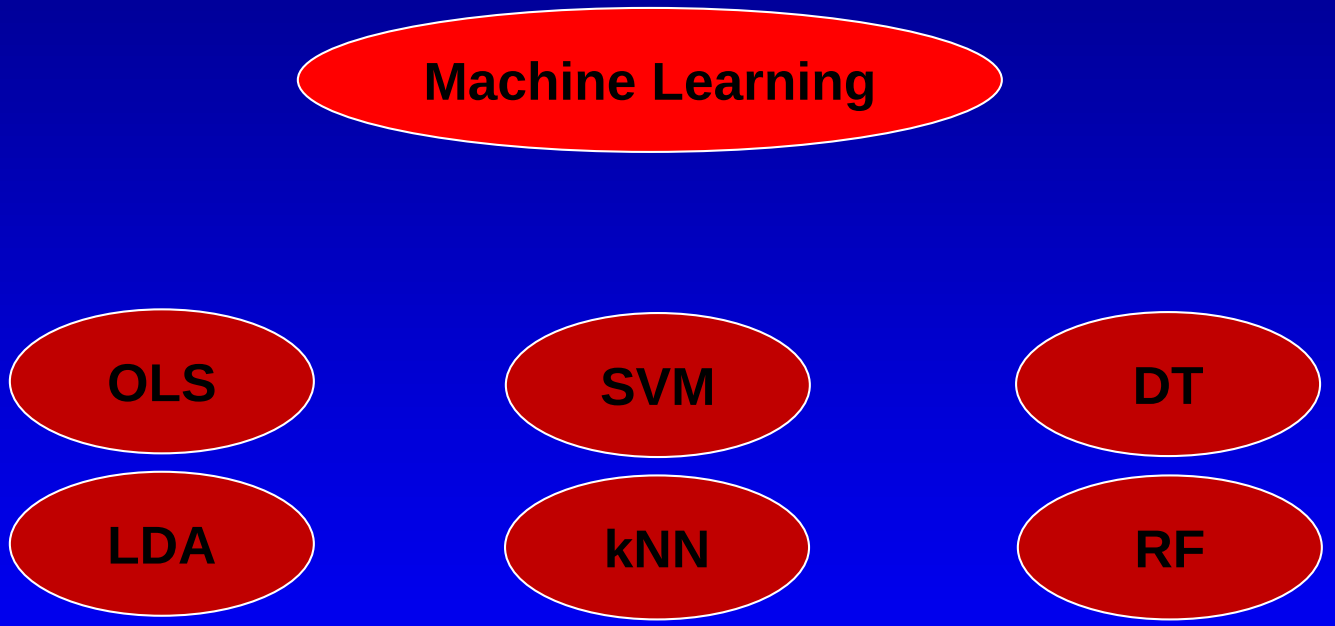
Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction. — Examples

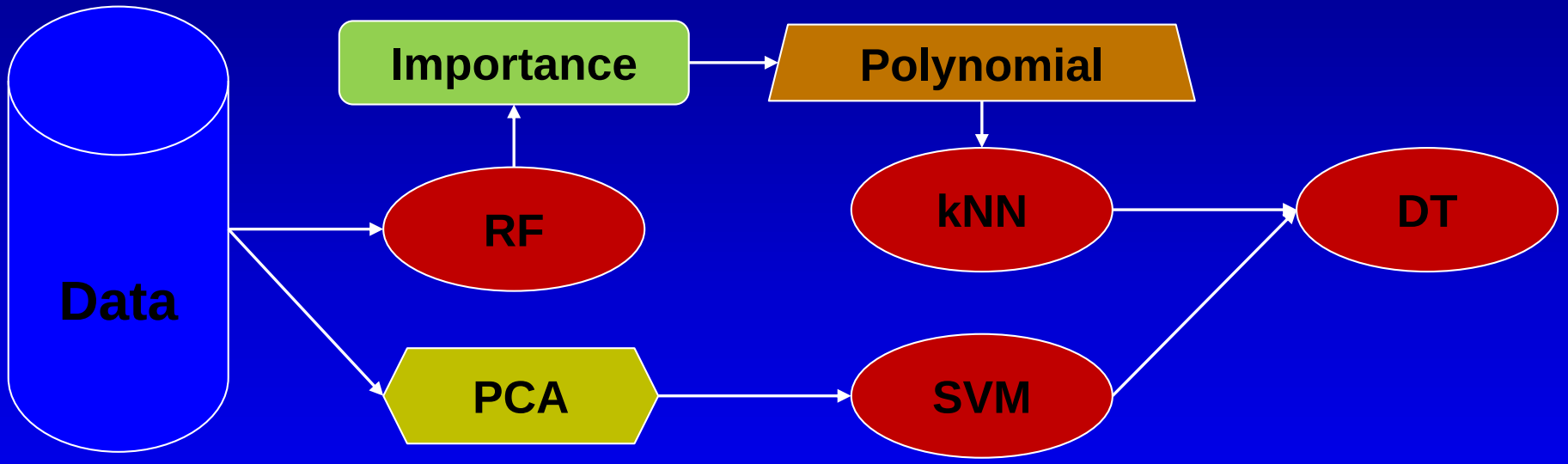
Building Blocks of a Pipeline



Building Blocks of a Pipeline

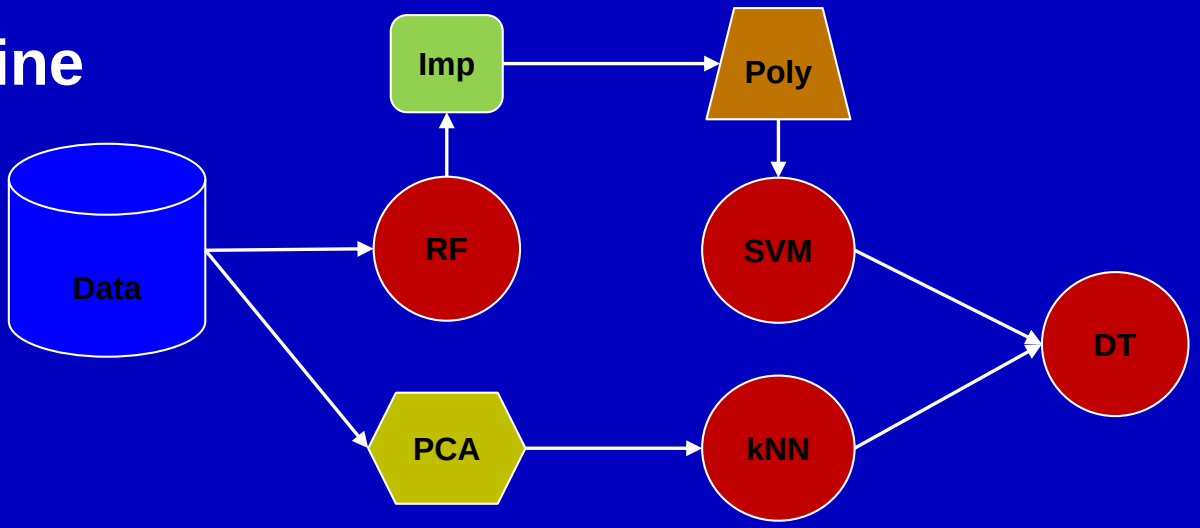


Example Data Science Pipeline

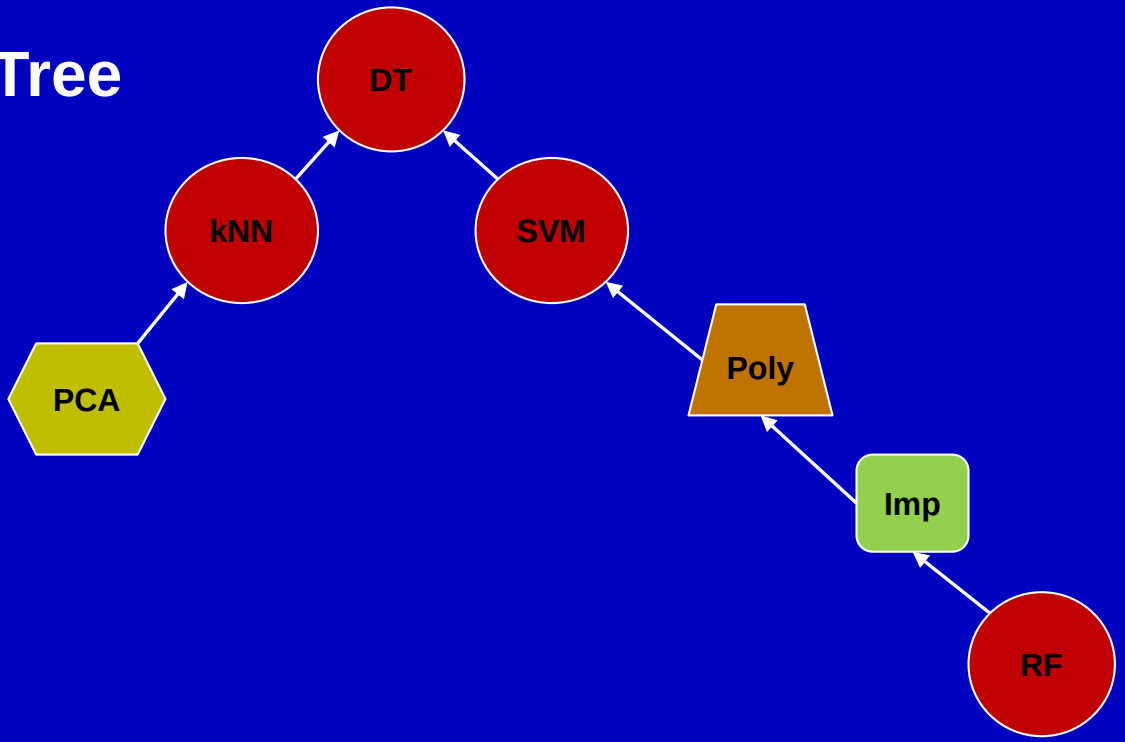


Pipeline representation

TPOT Pipeline

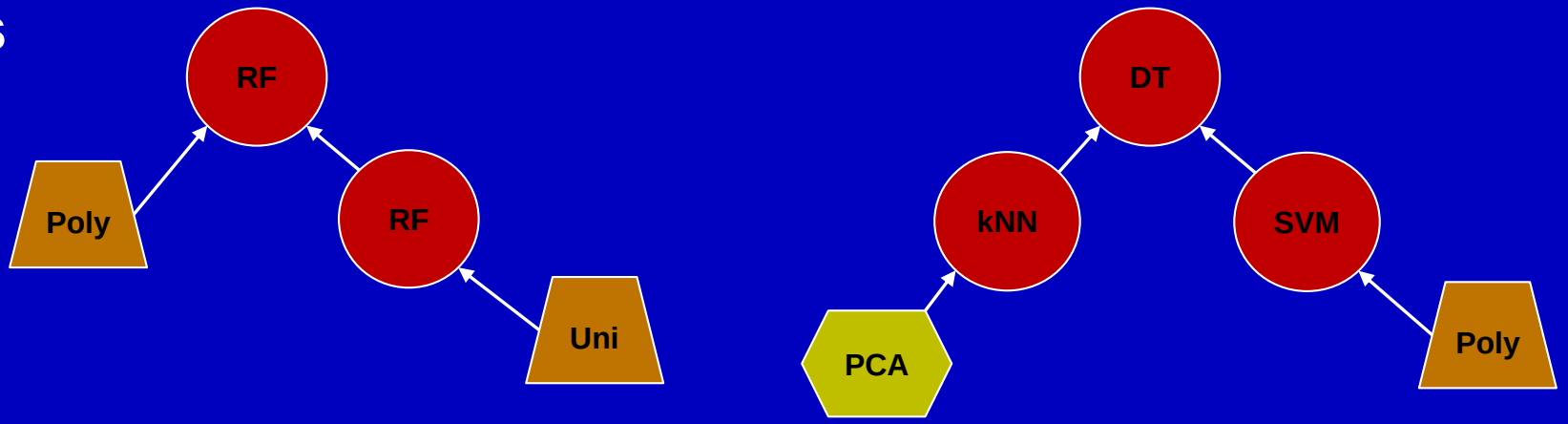


TPOT Expression Tree



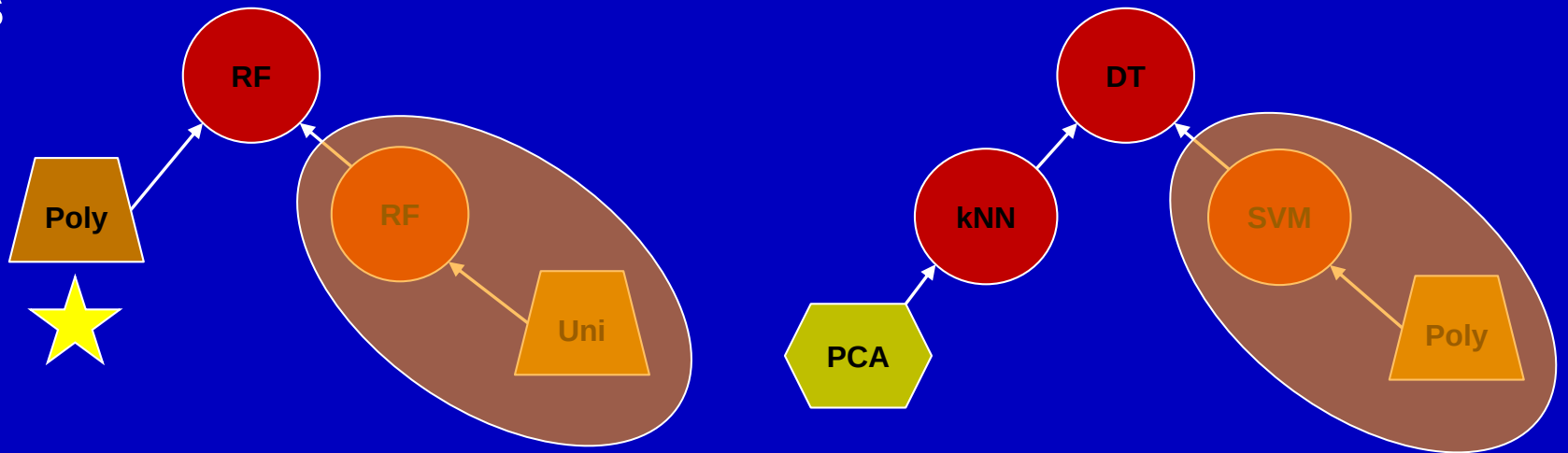
Pipeline optimization

Parents

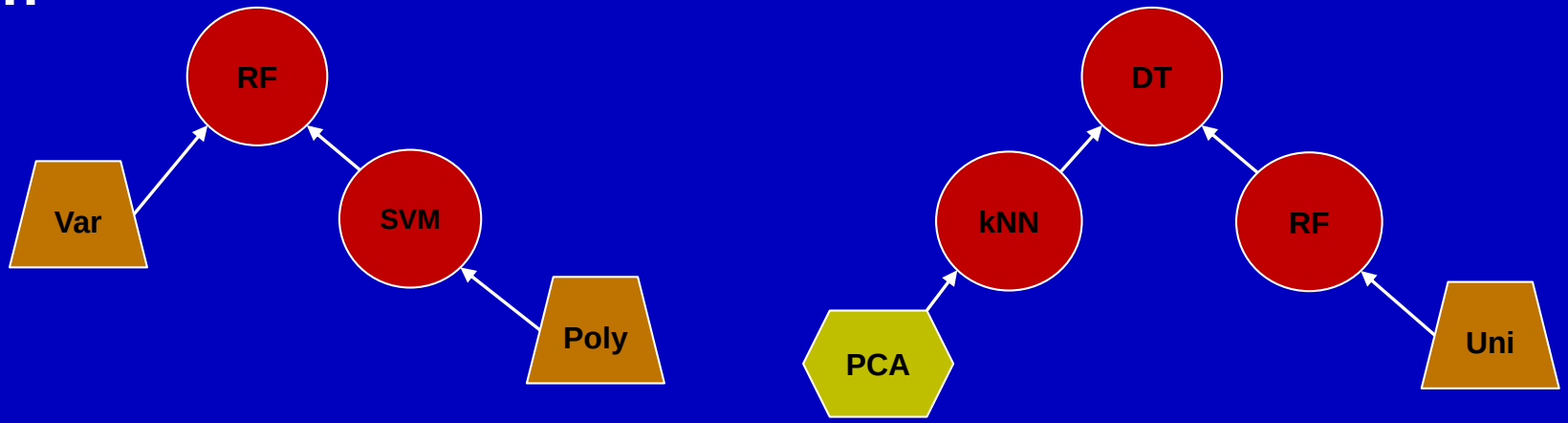


Children

Parents

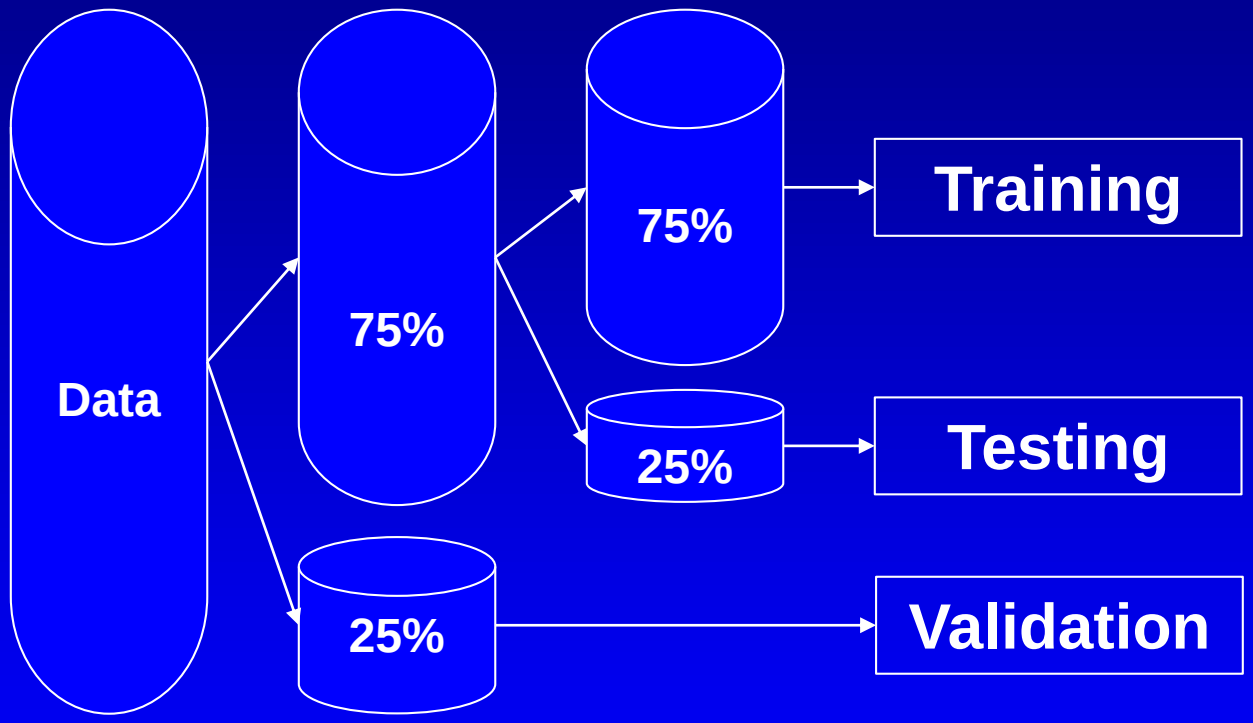


Children

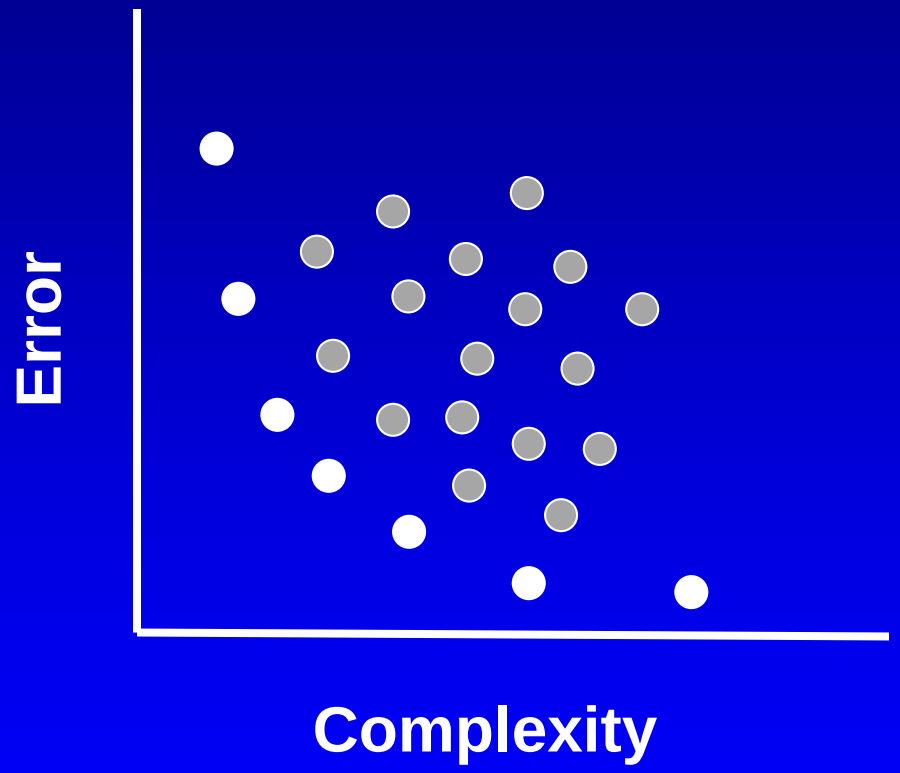


Pipeline overfitting

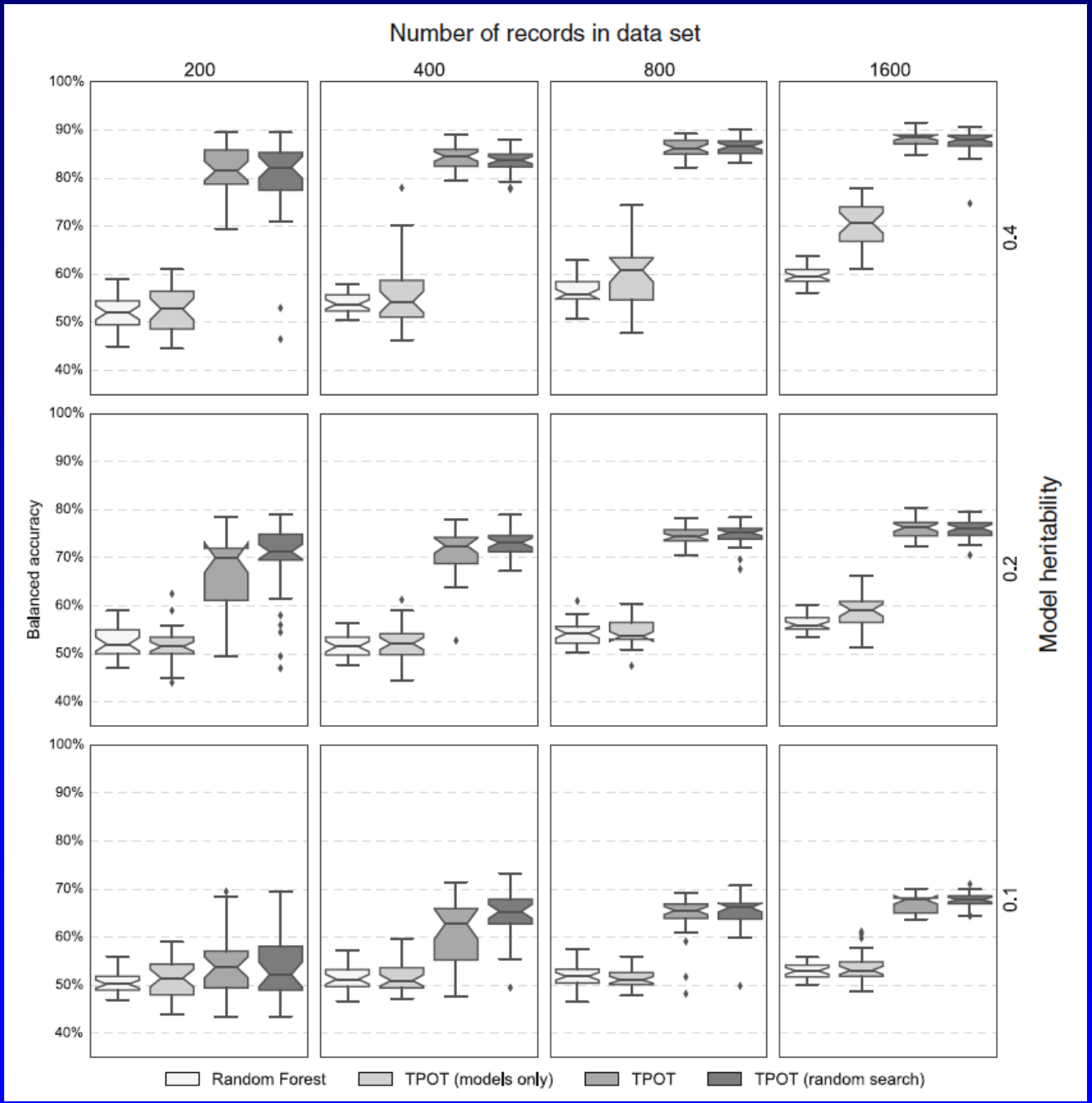
Cross-Validation



Pareto Optimization



Simulation Example



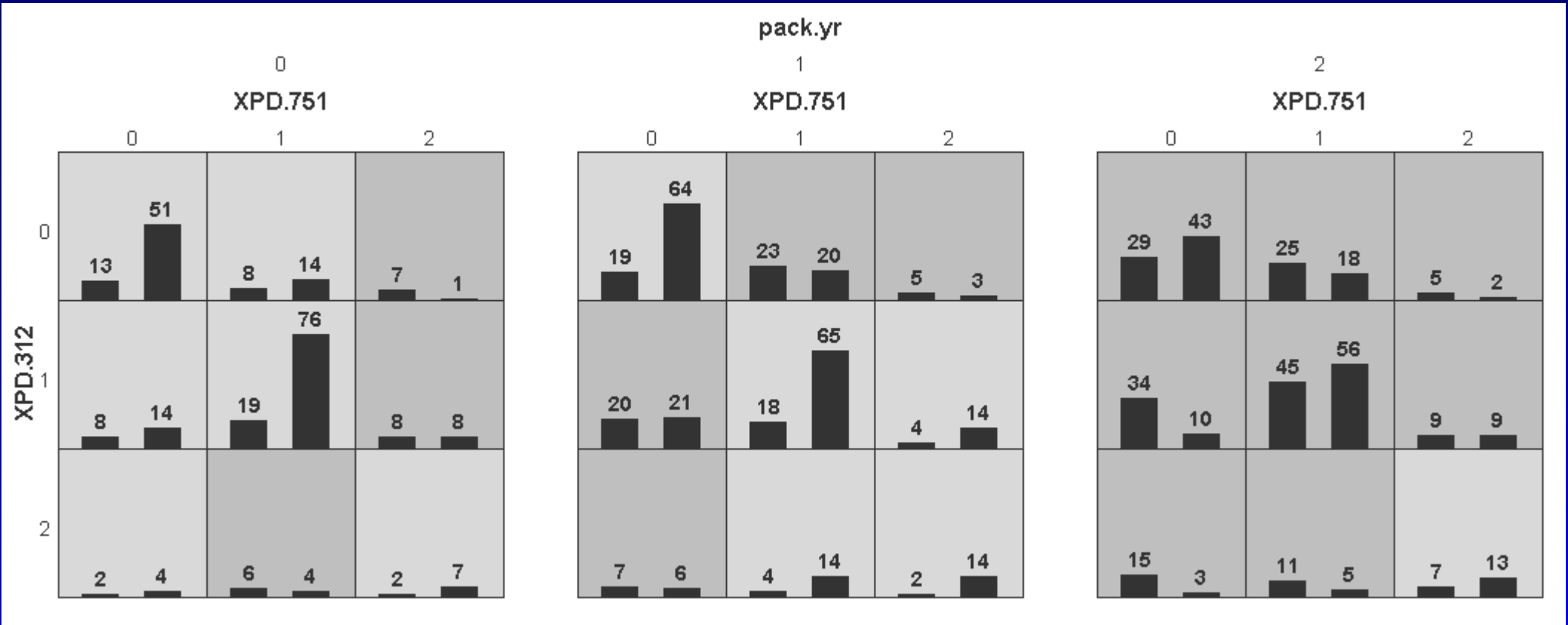
Olson et al., LNCS (2016)

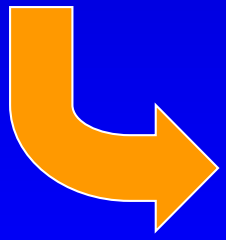
Bladder Cancer Example

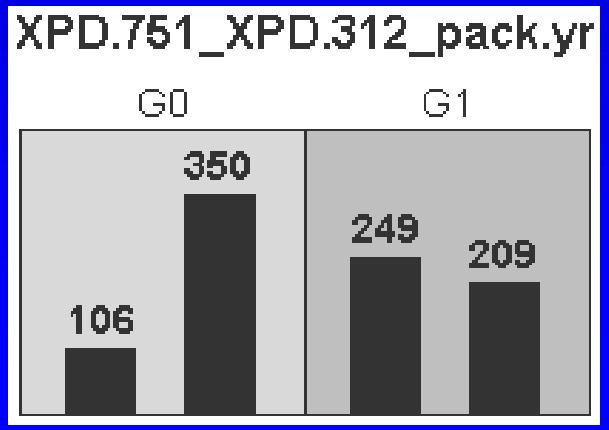
Application to Bladder Cancer

Andrew et al., *Carcinogenesis* (2006, 2008)

- 334 bladder cancer cases
- 580 controls
- From the state of NH
- Polymorphisms in DNA repair enzyme genes
  - *XPD*
  - *APE*
  - *XPC*
  - *XRCC1*
  - *XRCC3*
- Pack-years of smoking, age, gender







Training Accuracy 0.66
Testing Accuracy 0.64
OR = 3 (95% CI 2.0-3.4)
P < 0.001

TPOT Building Blocks

Feature Selection

- Variance
- Univariate
- Recursive
- Importance

Feature Processing

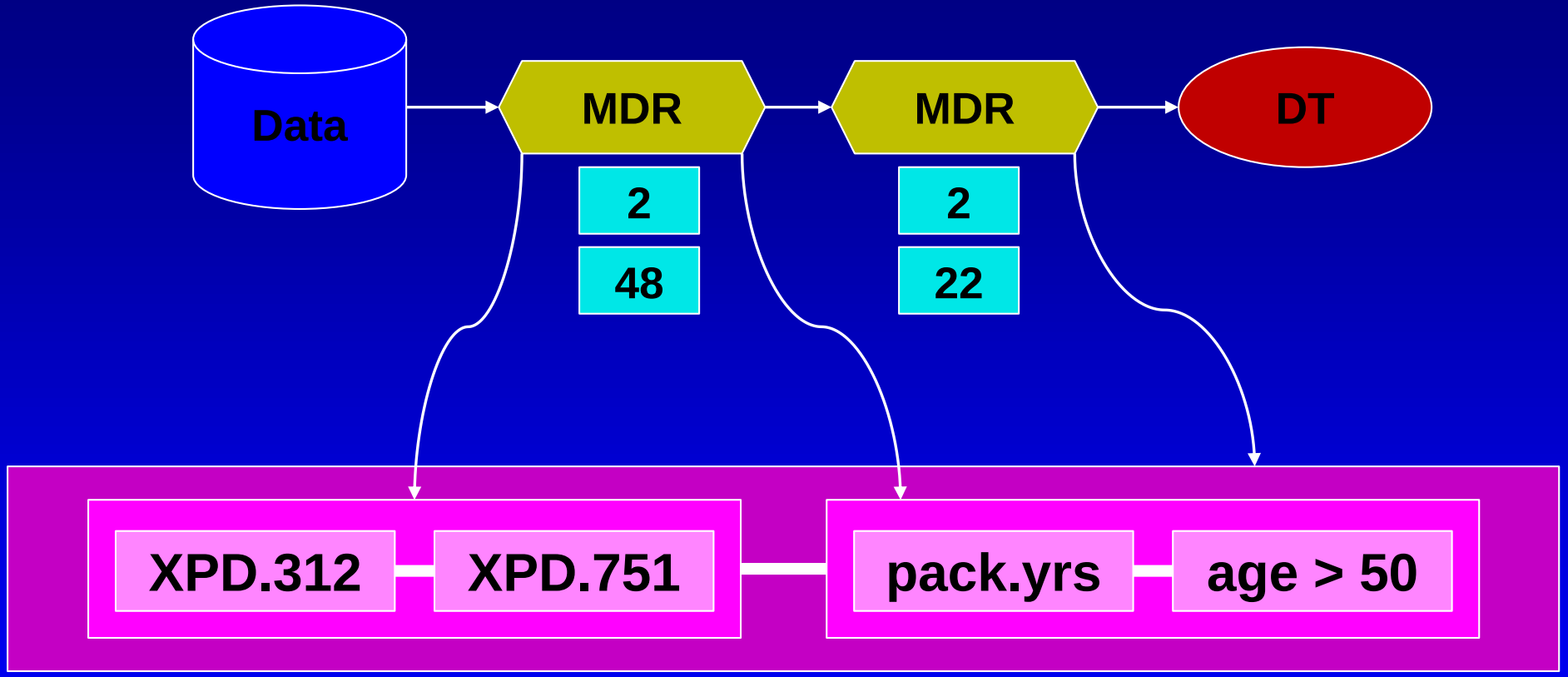
- Normalization
- Encoding
- Polynomial
- Scaling

Feature Construction

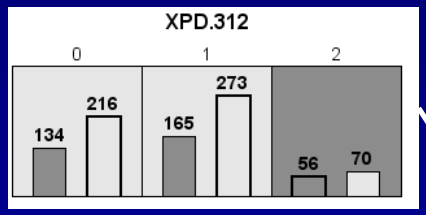
- PCA
- SVD
- Factor
- ICA
- MDR

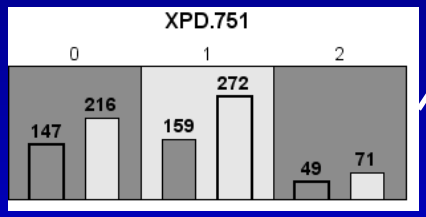
Machine Learning

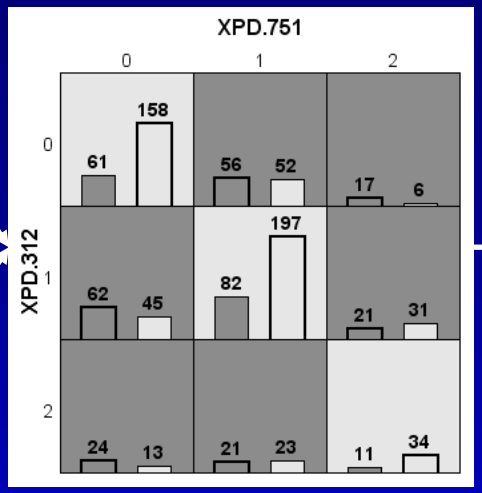
- DT
- NB

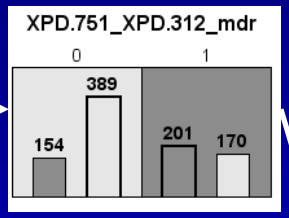


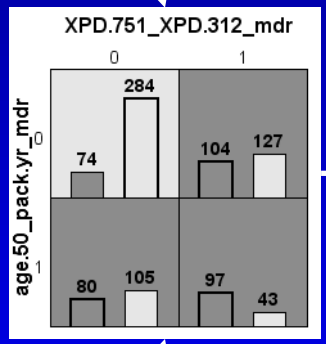
Testing Accuracy 0.64

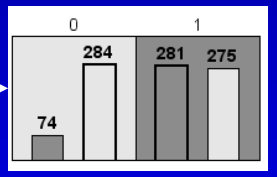

XPD.312
0
1
2
134
216
165
273
56
70

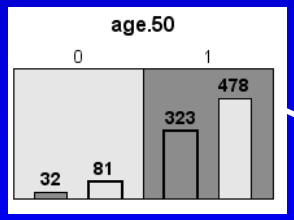

XPD.751
0
1
2
147
216
159
272
49
71

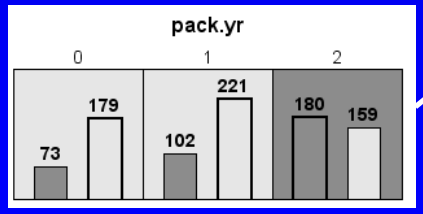

XPD.751
XPD.312
0
1
2
61
158
56
52
17
6
62
45
82
197
21
31
24
13
21
23
11
34

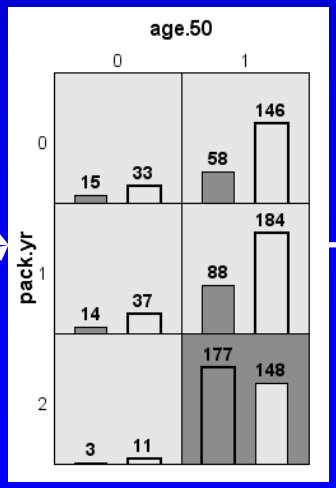

XPD.751_XPD.312_mdr
0
1
154
389
201
170

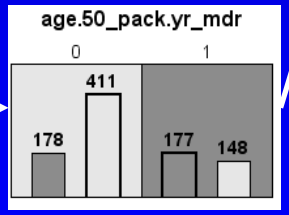

XPD.751_XPD.312_mdr
age.50_pack.yr_mdr
0
1
74
284
104
127
80
105
97
43

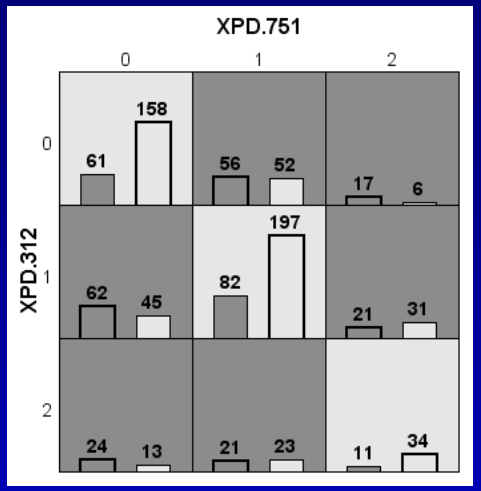

0
1
74
284
281
275

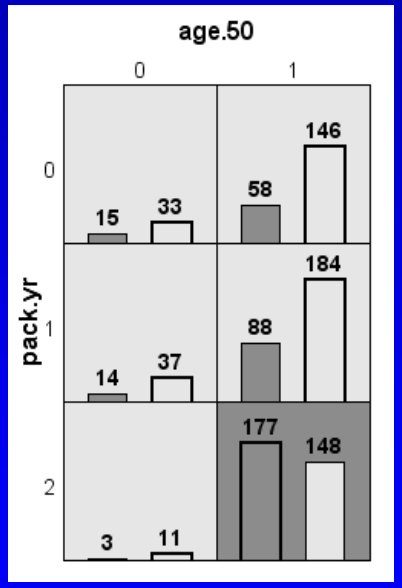

age.50
0
1
32
81
323
478

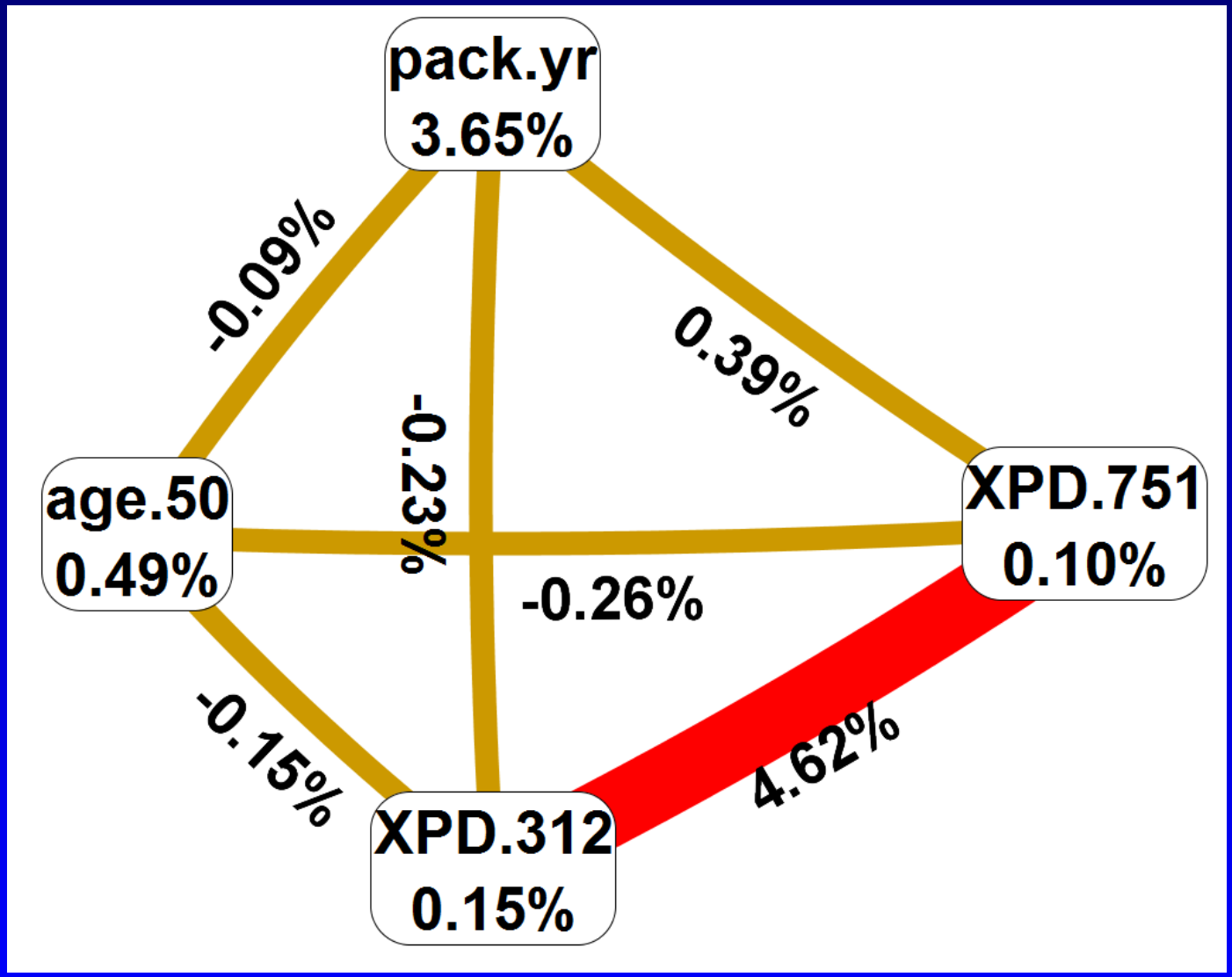

pack.yr
0
1
2
73
179
102
221
180
159


age.50
pack.yr
0
1
2
15
33
58
146
14
37
88
184
3
11
177
148


age.50_pack.yr_mdr
0
1
178
411
177
148

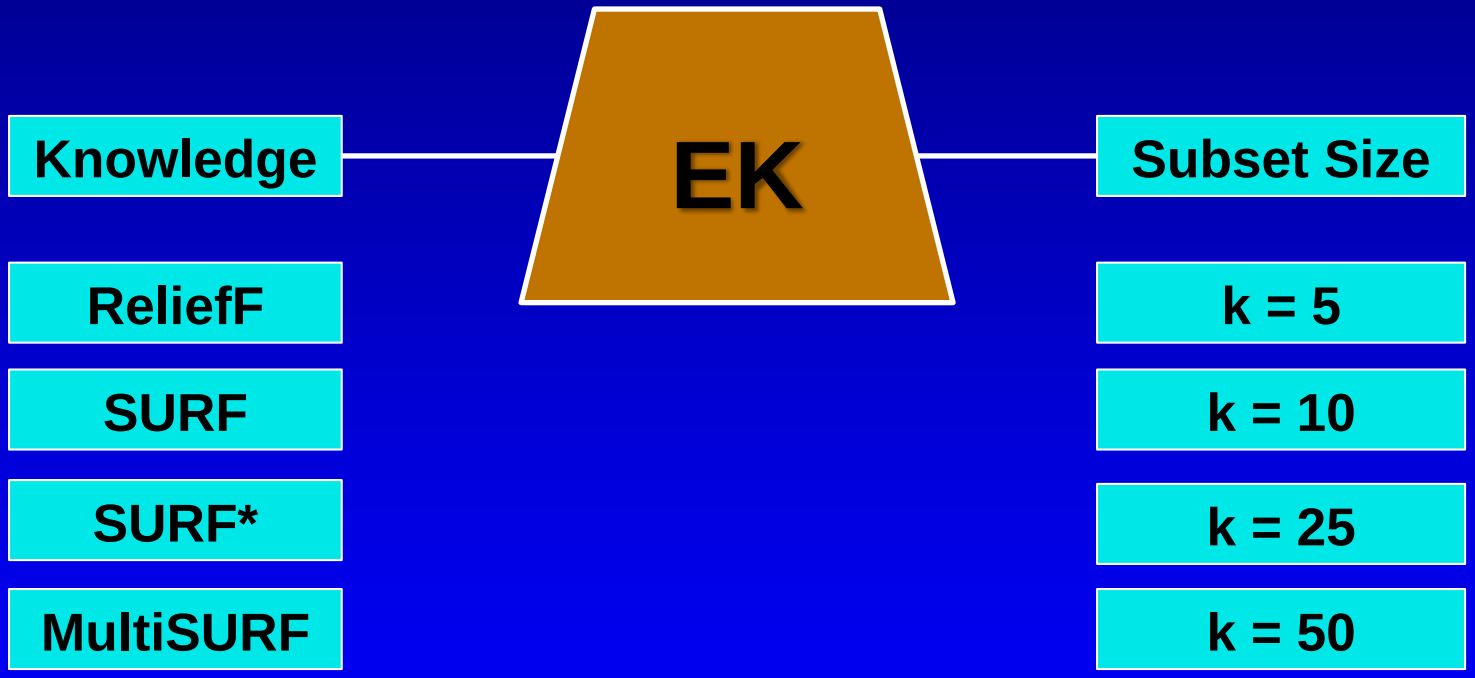

XPD.751
0
1
2
XPD.312
0
1
2
61
158
56
52
17
6
62
45
82
197
21
31
24
13
21
23
11
34


age.50
0
1
pack.yr
0
1
2
15
33
58
146
14
37
88
184
3
11
177
148

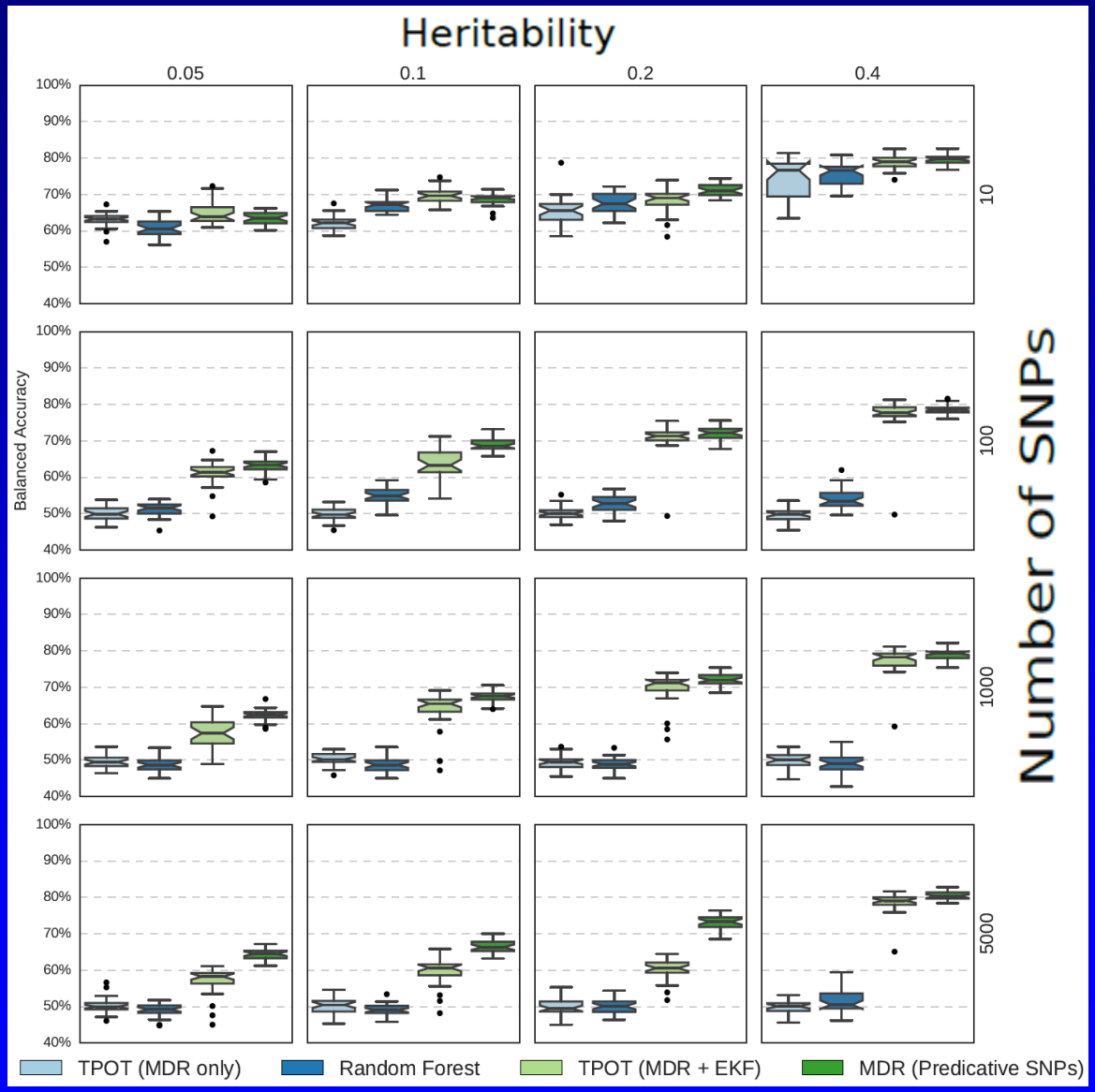

pack.yr
3.65%
age.50
0.49%
XPD.751
0.10%
XPD.312
0.15%
-0.09%
0.39%
-0.23%
-0.26%
-0.15%
4.62%

Scaling for big data

An expert knowledge feature selector



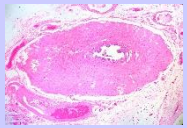
An expert knowledge feature selector

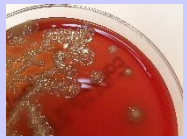


Sohn et al., LNCS (2017)

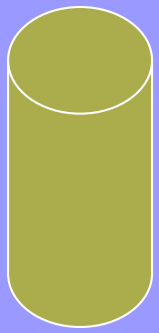
Automated Machine Learning (AutoML)

D





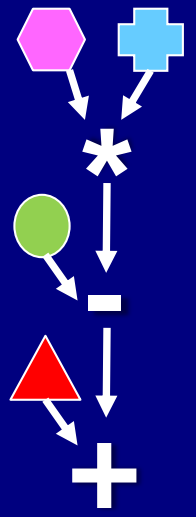
DI

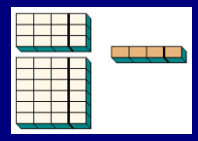


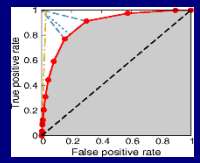
FS

FC

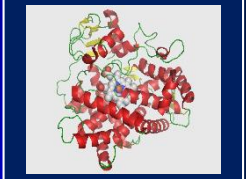
ML

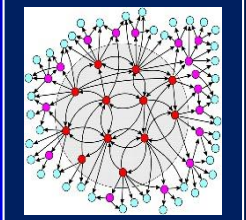


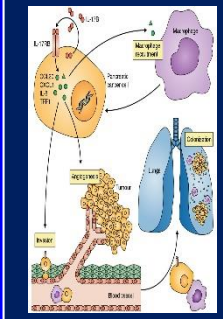




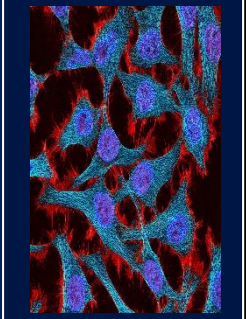
I



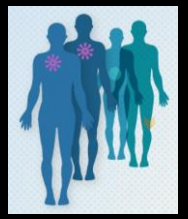




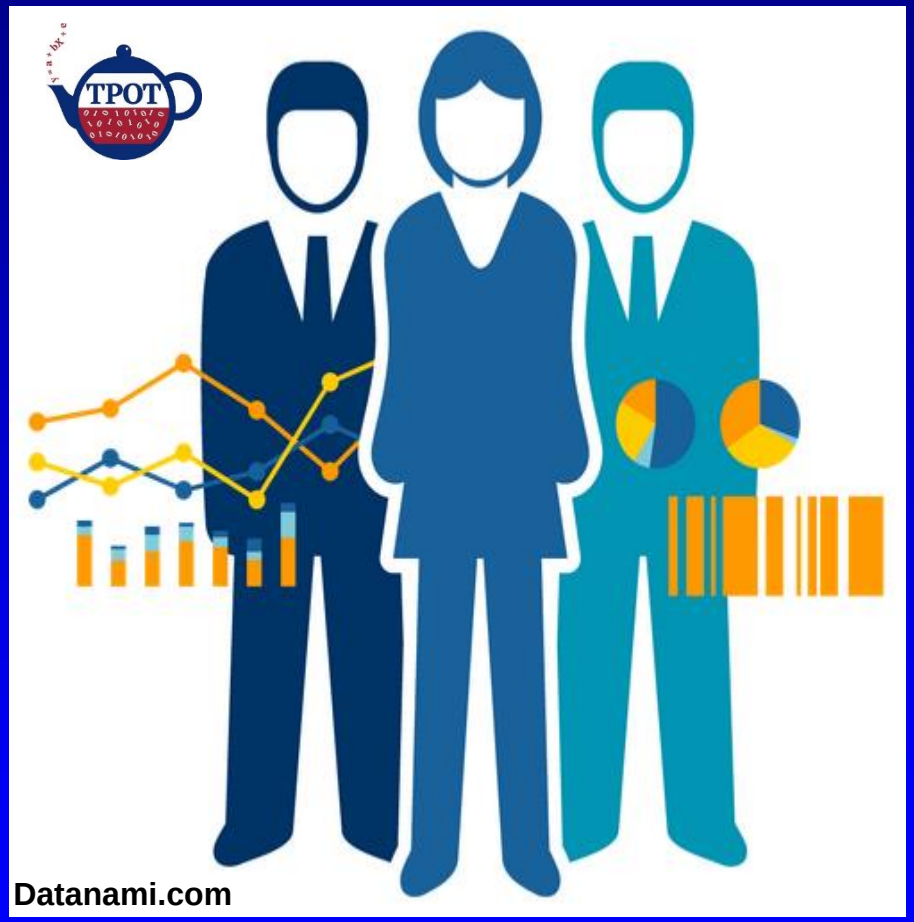
V

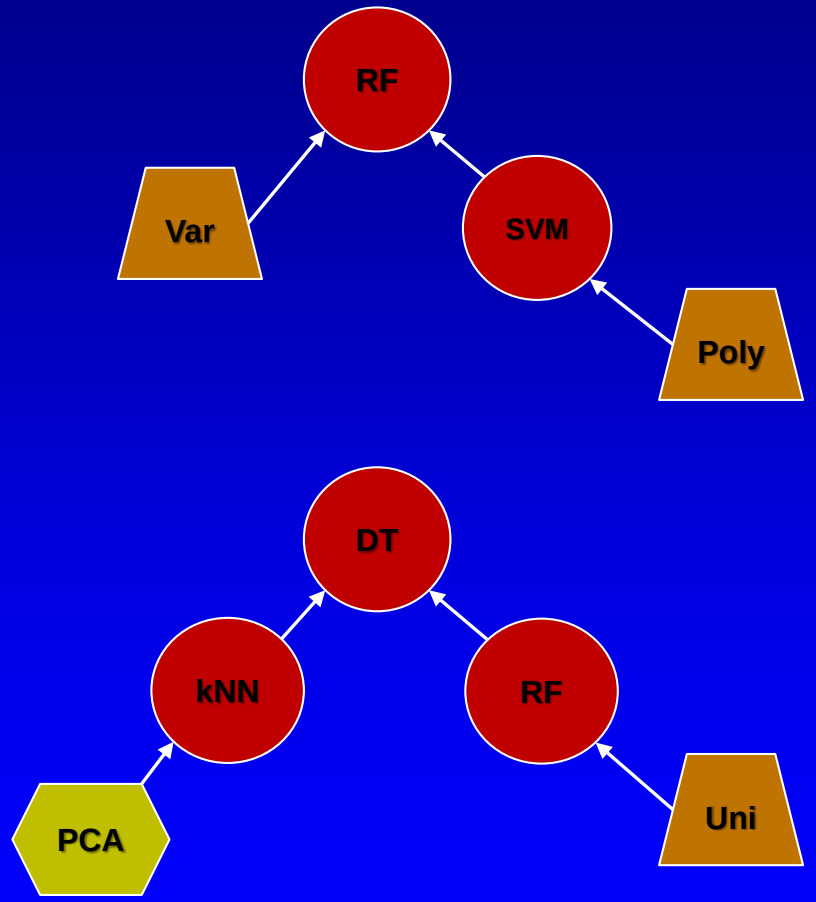


A

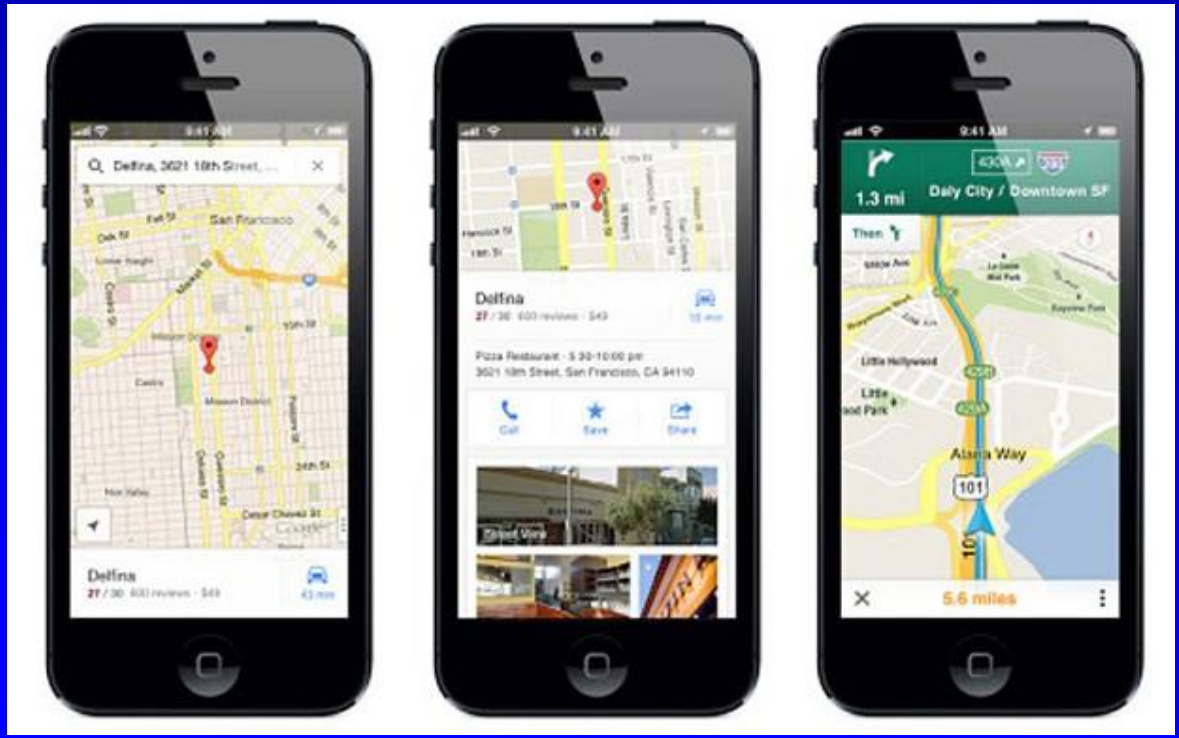


TPOT: The Data Science Assistant





Data Science should be *accessible* and *easy*



automl.info

AutoML

Information about Automated Machine Learning

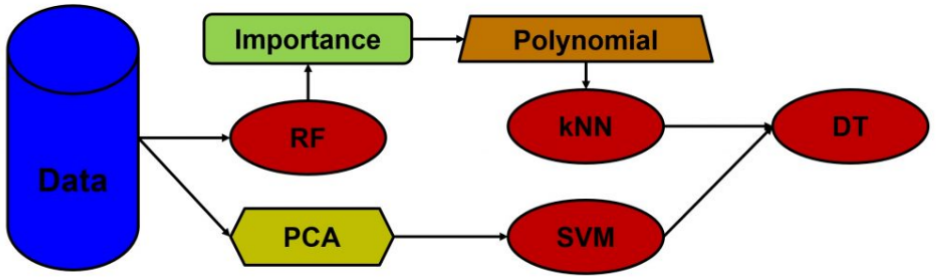
Home | AutoML | PennAI | TPOT | About Us

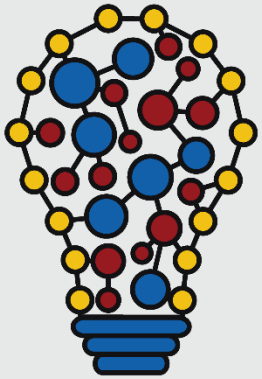
Information about Automated Machine Learning (AutoML)

The purpose of this site is to provide general information about the hot new field of automated machine learning (AutoML) and to provide links to our own PennAI accessible artificial intelligence system and Tree-Based Pipeline Optimization Tool (TPOT) algorithm and software for AutoML using Python and the scikit-learn machine learning library. We also provide links to some other commonly used AutoML methods and software.

The goal of AutoML is to make machine learning more accessible by automatically generating a data analysis pipeline that can include data pre-processing, feature selection, and feature engineering methods along with machine learning methods and parameter settings that are optimized for your data. Each of these steps can be time-consuming for the machine learning expert and can be debilitating for the novice. These methods enable data science using machine learning thus making this powerful technology more widely accessible for those hoping to make use of big data.

Below is an example of a hypothetical machine learning pipeline that could be discovered using a method such as TPOT. Here, the data are analyzed using a random forest (RF) with feature selection performed using the importance scores. The selected features then undergo a polynomial transformation before being analyzed using k nearest neighbors (kNN). The predictions made by kNN are then treated as a new engineered feature and passed to a decision tree (DT). In parallel, the data are also engineered using principal components analysis (PCA). The principal components are then passed as new features to a support vector machine (SVM) whose output is passed as an engineered feature to the DT with the other engineered feature. The DT then makes a final prediction. Each of the methods in this pipeline are included in the scikit-learn library.



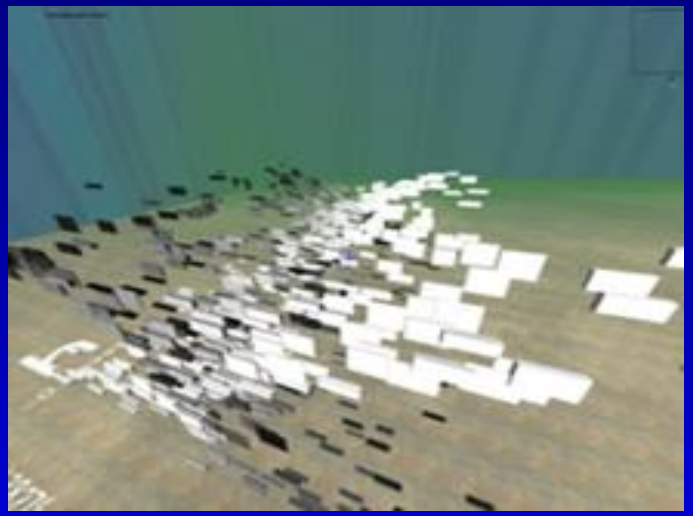


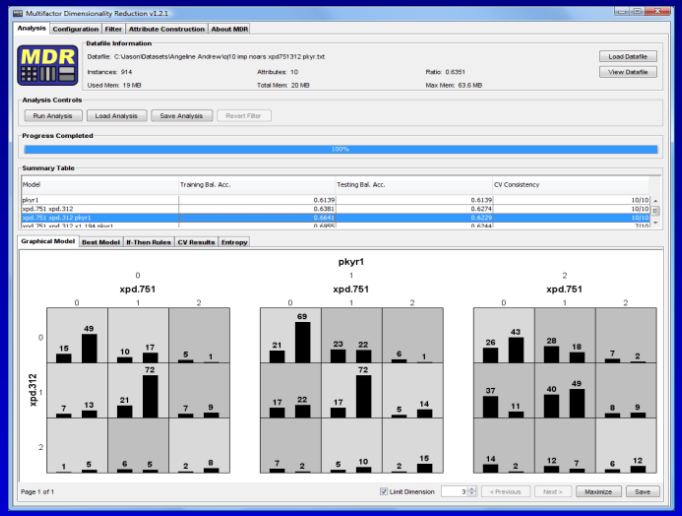
Penn Institute for Biomedical Informatics

Idea Factory

Penn IBI Idea Factory

Connecting Researchers with Ideas






Penn Institute for
Biomedical Informatics
Idea Factory

Acknowledgments

- Graduate Students and Postdocs
  - Brett Beaulieu-Jones, Brian Cole, Molly Hall, Bill LaCava, **Randy Olson**, Alena Orlenko, Patryk Orzechowski, Nadia Penrod, Elizabeth Piette, **Andrew Sohn**, Ryan Urbanowicz
- Staff
  - Elisabetta Manduchi, Pete Schmitt
- NIH grants R01 AI11679, R01 LM009012, R01 LM010098

- jhmoore@upenn.edu
- epistasis.org, epistasisblog.org
- twitter.com: @moorejh